STUDIES IN SPORT, PHYSICAL EDUCATION AND HEALTH

81

Liisa Heinilä

Analysis of Interaction Processes in Physical Education

Development of an Observation Instrument, and its Application to Teacher Training and Program Evaluation

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2002

STUDIES IN SPORT, PHYSICAL EDUCATION AND HEALTH 81

Liisa Heinilä

Analysis of Interaction Processes in Physical Education

Development of an Observation Instrument, and its Application to Teacher Training and Program Evaluation

Esitetään Jyväskylän yliopiston liikunta- ja terveystieteiden tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston vanhassa juhlasalissa (S212)
helmikuun 23. päivänä 2002 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Sport and Health Sciences of the University of Jyväskylä,
in Auditorium S212, on February 23, 2002 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2002

Analysis of Interaction Processes in Physical Education

Development of an Observation Instrument, and its Application to Teacher Training and Program Evaluation

STUDIES IN SPORT, PHYSICAL EDUCATION AND HEALTH 81

Liisa Heinilä

Analysis of Interaction Processes in Physical Education

Development of an Observation Instrument, and its Application to Teacher Training and Program Evaluation

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2002

Editors
Harri Suominen
Department of Health Sciences, University of Jyväskylä
Pekka Olsbo and Marja-Leena Tynkkynen
Publishing Unit, University Library of Jyväskylä

URN:ISBN:978-951-39-9123-4
ISBN 978-951-39-9123-4 (PDF)
ISSN 0356-1070

Jyväskylän yliopisto, 2022

ISBN 951-39-1118-7
ISSN 0356-1070

Jyväskylä University Printing House, Jyväskylä
and ER-Paino Ky, Lievestuore 2002

ABSTRACT

Heinilä, Liisa
Analysis of interaction processes in physical education. Development of an observation instrument, its application to teacher training and program evaluation. Jyväskylä: University of Jyväskylä, 2002, 406 p.
(Studies in Sport, Physical Education and Health
ISSN 0356-1070; 81)
ISBN 951-39-1118-7
Yhteenveto
Diss.

This dissertation deals with a large-scale and long-term research, which focussed on the measurement of teaching process – interaction – in physical education. The main aim was to contribute to strengthening the knowledge base for developing physical education. The framework of the research strategy used in this research consisting of three phases, and the results obtained in meta-level and substantive level as assessment of reliability and validity of the developed measuring instruments, is presented in the first and second part of the dissertation

The final phase focussed on Flanders-based (Flanders 1965, 1970, Heinilä 1974, 1977b) preservice study unit program (80 h), more specifically on the predictive validation of the course of didactic observation and microteaching in the frame of the contextual setting, before and after the study degree program reform at the faculty (1978), by examining variation of predictability of students' study success. Based on theories and assumptions of (i) the learning process that produces particular teaching behaviors (eg. non-directive teaching) and (ii) on the basis of the study of the relationships between factors related to context, presage, content, process and outcome, a model was constructed (and tested), which assumes that the students' study success can be predicted by its relationship with students' background characteristics assessed during the selection procedure: the sum score in stage 1 including eg. prior school success, stage 2 measures consisting of the theory test, practice skills test, entry teaching behavior (teaching episode), and teaching behavior (control data) measured two years later (PEIAC/LH-75 II, Heinilä 1977b) as well attitude measures. Using replicated designs and hierarchical regressions analyses with the male, female and total groups of four course groups (n = 205), it was possible to confirm assumptions about the learning process and about the relationships between the course, students' background factors and the outcomes of the course processes, as is shown eg. by the results of the case study (1988, n = 42) carried out after the study degree program reform. In the obtained prediction model (R^2 = .35, F(4,37) = 4.86, p = .003), the intake tests accounted for a major part of the explained variance (57 %) of study success (final grade) due to the combined effects of the theory test scores (12 %), intake stage 1 sum scores (8 %) as well as their interaction effects with students' attitudes (7 %) and entry teaching behavior (7 %). Additionally, the assumption of the relationship between gender and study success (determined already in the selection stage) was supported, whereas the assumption of the predictive power of the practical skill tests did not receive support. However, the internal and external validity of the didactic observation and microteaching course program was supported by students' (n = 283) evaluation of the course.

Key words: teaching behavior, physical education, interaction process analysis, observation instrument, didactic observation and microteaching, program predictive validation.

Author's address

Liisa Heinilä, Lic.
Department of Physical Education
University of Jyväskylä, Finland

Supervisor

Professor Lauri Laakso
Department of Physical Education
University of Jyväskylä, Finland

Reviewers

Professor Udo Hanke
Koblenz-Landau University, Germany

Professor Jouko Kari
Department of Teachers Education
University of Jyväskylä, Finland

Opponents

Professor Udo Hanke
Koblenz-Landau University, Germany

Professor (emeritus) Sauli Takala
Faculty of Humanities
University of Jyväskylä, Finland

Dedication

This work is dedicated to the memory of my mother, Laila, Marja Saarenmaa, Medical Counsellor, Doctor of Dentistry who loved science and hard work.

ACKNOWLEDGEMENTS

I would like to express my gratitude to a number of individuals who have contributed to the different phases of my work over the past thirty years. The conception and eventual implementation of this study is the result of my interaction with several members of the Institute of Education, University of Helsinki, and of the Faculty of Sport and Health Sciences, University of Jyväskylä. The fact that I can mention only a few does not diminish my recognition of their contributions to my thinking and planning during this project. The last report of the project (Heinilä 1992a) was classified into the category "reflective teaching" by Claude Paré 1995 (ed.) in Proceedings of the AIESEP International Seminar on Training of teachers in reflective practice in Physical education "Better Teaching in Physical Education? Think about it!" - which motivated me to complete this life-long process - this dissertation.

I wish to express a special note of appreciation to Professor Matti Koskenniemi, my teacher at the University of Helsinki, who, in 1967, initiated and directed the research project called the Instructional Process Analysis in the Nordic Countries (DPA-Project) and Professor Erkki Komulainen, a member of this research group. I have received inspiration from his extensive scientific production. His critical reading of early drafts and his significant remarks and suggestions are very greatly appreciated. My thanks also due to Professor Pentti Pitkänen who helped me to prepare for my general examination at the Faculty of Sport and Health Sciences, University of Jyväskylä. He has encouraged and helped me to continue this research project and to apply its results to the design and implementation of teacher training programs in the early 1970s. I also thank Professor Risto Telama and Professor Lauri Laakso who have continued in the institutional context to support this study project. Especially, I will thank Professor Taru Lintunen and Professor Pilvikki Heikinaro-Johansson, who worked so diligently during the training sessions, observation periods and later motivated, and gave their critical reflections and support to complete this large-scale study. Professor Sauli Takala has been especially helpful in suggesting how to recast the outline and content of many years' work into a single and coherent English language text.

I also express warm thanks to my reviewers professor Udo Hanke and professor Jouko Kari, for their constructive comments on the manuscript. Without the involvement of teachers and student teachers in creative ways, the plan could not have been realized. I am very grateful to the research assistants Olli Akkanen, Lasse Hellsten, Sanna Klemola, Merja Luuppala, Kirsti Partanen, Elina Ora, Päivi Paukku, Anita Pälvimäki, Pekka Reponen, Tapio Tammivuori and Jari Utriainen.

Many thanks also are due to Mauno Väisänen, Pekka Sipilä and Jukka-Pekka Kesonen, who performed the statistical analyses at the University of Jyväskylä computer center and to Reijo Pirttimäki who with his assistants recorded and videotaped the material. Previous reports of this study were translated or revised by Glyn Hughes, Douglas Robinson, Dr. Sauli Takala, and

Ilkka Rekiaro. Thanks also to Dr. R. Elaine Degenhart for her contribution in editing and correcting the first section English language manuscript.

I am indebted to Liisa Nieminen, Lahja Kilpeläinen, Ritva Tanabe, Raili Puranen, Taru Venäläinen, Minna Jokinen and Mervi Venäläinen for their exellent secretarial work.

And, finally, I will thank my own family for the patience shown for my engagement in this longterm research work.

This study was made possible by the financial support of the Ministry of Education, the Ellen and Artturi Nyyssönen Foundation, the Finnish Cultural Foundation and the University of Jyväskylä.

Jyväskylä, February 2002

Liisa Heinilä

PREFACE

The research program reported here consisted of phases, which have been reported separately within a period of 30 years (1970-1997).

The purpose of this dissertation is to do the following:

1) create a synthesis of the reports related to the study program
2) report the basis of decisions made in constructing the observation instruments PEIAC-LH/75 and PEIAC-LH-75 (II)
3) present the created measuring instrument
4) report the explorative studies made for determining the capacity of the proposed instrument for gathering and organizing data in physical education teaching events based on a framework developed after surveying relevant research literature,
5) report the investigations for the development of a microteaching program package and for determining its internal and external predictive validity
6) discuss the results both from development and application perspectives

This dissertation is based on the following original articles, and technical reports, which will be referred to in the text:

Heinilä, L. 1970. Opettajan ja oppilaiden välisistä vuorovaikutussuhteista liikunnan opetustilanteissa. (About teacher-pupil interaction in physical education classes). Reports of the Finnish Society for Research in Sports and Physical Education no. 22. Helsinki: The Finnish Society for Research in Sport and Physical Education, 80-94. ISSN 0561-7731, UDK 796/799.

Heinilä, L. 1971. Liikunnan opetustapahtuma sosiaalisena vuorovaikutus-prosessina (Teaching of physical education as a process of social interaction). University of Jyväskylä, Finland. Unpublished master's thesis.

Heinilä, L. 1974. Developing a system for describing teacher-pupil interaction in physical education classes. Paper presented at FIEP scientific congress Gdansk 27-31. May 1974. In T. Bober and G. Mlodzikowski (eds.). Education physique des enfants avant l'Epoque de la Puberte. Edition Scientifiques de Pologne, Warsaw, Monographie no 12, Gdansk 1976, 218-223, and FIEP Bulletin 1974, 44(4), 16-20 (Eng.), 59-62 (French).

Heinilä, L. 1976. Objectivity of coding in a system (PEIAC/LH-75) developed for describing teacher-pupil interaction process in physical education classes. In T. Haajanen and M. Veistola (eds.) Research in Physical Culture in Finland, Policy in Physical Culture Research Work, Abstracts IV 1976. Reports of the Finnish Society for Research in Sports and Physical Education, 1977, 55 and 49. Helsinki: Finnish Society for Research in Sports and Physical Education, 66, 22-23.

Heinilä, L. 1977a. Analysing systems in the evaluation of the teacher-pupil interaction process in physical education classes. In Tammivuori (ed.), Evaluation: International Congress of Physical Education, July 1976, University of Jyväskylä. Congress proceedings of the Finnish Society for Research of Physical Education and Sport no. 64. Helsinki, 1979, 37-58. FIEP Bulletin 1977, 47(1), 20-34 (Eng.), 47(1) 13-25 (French). FIEP Bulletin 1978 48(3), 4-23 (Portug.). Methode d' evaluation du processus d'enseignemet en education physique, FFGEV-Gymnastique. Volontaire 1, 1977, 24-33 (French).

Heinilä, L. 1977b. Application of interaction analysis to the teacher education in physical education. Paper presented at the International AIESEP-FIEP Congress of Physical education and Sports, Madrid June, 1977. Research reports from the Departement of Physical Education, University of Jyväskylä, 15. (1979) and Research Bi-Annual for movement. Manhattan-State India 13 (2) 1997, 16-56.

Heinilä, L. 1980. Developing a system (PEIAC/LH-75) for describing teacher-pupil interaction in physical education classes: Objectivity and content validity of coding. Paper presented at the International AIESEP-congress in Magglingen 10.-16.9.1978. In G. Schilling & W. Bauer (eds.), Audiovisual Means in Sport. Basel: Birkhaus Verlag, 361-370.

Heinilä, L. 1983. Developing a system (PEIAC/LH-75) for describing teacher-pupil interaction in physical education classes: Construct validity and sensitivity. In R. Telama, V. Varstala, J. Tiainen, L. Laakso & T. Haajanen (eds.), Research in school physical education. AIESEP congress 1982 Jyväskylä. Finland. Reports of the Foundation for Promotion of Physical Culture and Health, 38, 124-132.

Heinilä, L. 1987. The development, validitation and application to teacher training of a system (PEIAC/LH-75) designed to expand the Flanders system of interaction analysis for describing teacher-pupil interaction process in Physical Education classes, published lic. thesis University of Jyväskylä; 20.10.1987.

Heinilä, L. 1988 Selecting students for physical education teacher education programmes. FIEP. Bulletin 58, 2/3 1988, 29-42.

Heinilä, L. 1990. Validation of an observation system in physical education: A multivariate approach. Research report presented at the International AIESEP Congress, Trois-Riviers, Québec, Canada. 1987. In M. Lirette, C. Paré, J. Dessureault & M. Piéron (eds.) Intervention en Éducation Physique et en Entrainement, Bilan et Perspectives. Physical Education and Coaching, Present State and Outlook for the Future. Québec: Presses de l'Université du Québec, 28-40.

Heinilä, L. 1992a . Prediction of success in student teaching from students selection variables, rated and measured teaching behavior and students' attitudes. Research report presented at the Olympic Scientific Congress 14.-19.6.1992 in Malaga, Spain. Actas Congreso Cientifico Olympico 1992. Pedagogia y Education Fisica Comparada. Serie Deporte y Documentation Instituto Andaluz Del Deporte no. 24, 1995, vol. III, 54-62. Also in references C. Paré (ed.) Better teaching in Physical Education? Think about it! (1995) Canada, Trois-Rivièrs: University of Québec, 291.

TABLE OF CONTENTS

SECTION I

1 INTRODUCTION

A central task of the university is the planning, realization and evaluation of goal-directed educational programs. This activity should be long-term, comprehensive and integrated with general social planning. It should also be closely linked with decision-making concerning all education. The ultimate aim of educational planning should be the quantitative and qualitative development of education (Itälä 1969). The development of educational programs is a multistage process at several levels and should be based on scientific research.

Attempts were made early in the 20th century to apply the methods of scientific research to the problems of school learning, teacher behavior, and teacher education. Within the behavioral sciences there has emerged a sub-discipline of "research on teaching", which Gage (1972) has defined in the following way:

> "Research" is defined as scientific activity aimed at increasing our power to understand, predict, and control events of a given kind. All three of these goals involve relationships between variables. ... "Teaching" in turn may be defined as events, such as teacher behavior, intended to affect the learning of a student. ... Given these definitions of "research" and "teaching", we can define "research on teaching" as the study of relationships between variables, at least one of which refers to a characteristic or behavior of a teacher. If the relationship is one between teacher behaviors or characteristics, on the one hand and effects on students, on the other, then we have "research on teacher effects", in which the teacher behavior is an independent variable. If the teacher behavior or characteristic serves as a dependent variable in relation to some variable in the program of selecting and training teachers (the teacher education program), then we have "research on teacher education". Both kinds of research taken together make up the field of research on teaching. (pp. 16-17)

This definition does not suggest that other kinds of variables are not also useful, and in fact desirable, in research on teaching. It only specifies that the variables

of teacher behavior and characteristics are at the center of concern and must be involved. Figure l illustrates the relationships in Gage's definition.

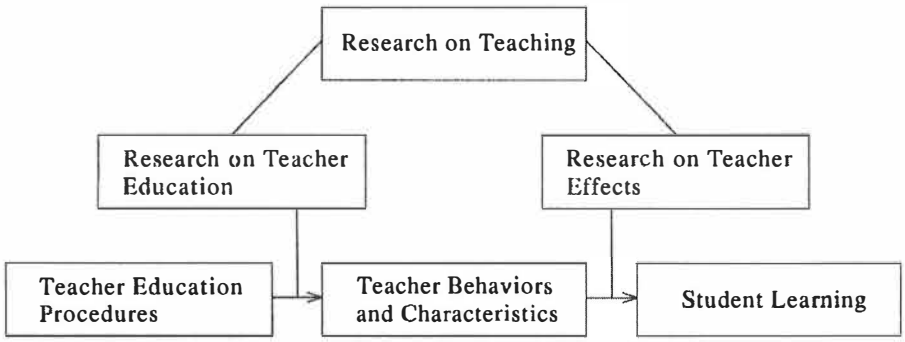


FIGURE 1 The field of research on teaching (Gage 1972, p. 17)

It has been suggested (Binet 1918) that everything has been said in education while nothing has been proved. It is true that much has been done since the early decades of the 20th century, but it is similarly true that several problems need to be addressed before practice teaching, and indeed, teacher training in general, can be fully developed. Only two of these problems will be taken up here. First, we need to have a feasible and comprehensive conceptualization of the nature of teaching. Second, we need reliable, valid and practicable ways of describing, analyzing and evaluating teaching activities and behaviors. Finally, having addressed these problems, we need to apply what we have found to teacher training programs.

1.1 Interaction analysis methods

The increasing emphasis on interaction and communication between teacher and students and among students, and the subsequent development of methods of interaction analysis has had a profound impact on empirical research on teaching. At an early stage of this new research paradigm, there was a clear interest in studying what contributions interaction analysis might be able to make to teacher education and practice teaching.

Interaction analysis is a label that refers to any technique for studying the chain of classroom events in such a fashion that each event is taken into consideration (Flanders 1970). The method is based on a conception of teaching as an interpersonal influence whose purpose is to affect pupil learning in line with set objectives. Typical of teacher behavior is human voice and motion, but it may also be frozen in the form of a book or film or a set of programmed

instructional materials (Gage 1972). In the study of teacher behavior, this influence can be observed on the basis of variable values placed on given dimensions such as teacher-centered/pupil-centered, direct/indirect, etc., and event sequences can be described, for instance, by means of a timeline display (cf. Flanders 1970).

Methods of interaction analysis are based on theoretical considerations and thus contain given conceptual systems. This is true also of the well-known systems developed by Bales (1950) and Flanders (1965, 1970). Thus, in using methods of this kind the researcher has not only made methodological decisions, but also he has bound himself to a particular theory and set of variables (Heinilä 1974, 1977a). In this way the measuring instrument achieves a central significance.

It is, therefore, hardly surprising that interaction analysis methods have also proved to be an effective tool in teacher training. They provide a conceptual scheme and simultaneously the means for the operationalization and measurement of variables. Perceptions and communications become more unified and precise, evaluation and comparison attain higher objectivity. The *contents* of teaching programs refers to the matter being dealt with, such as command words in practice teaching in P.E., or other forms of social interaction, different types of ball games, etc. *Form of teaching* refers here to the way in which interpersonal communication is organized (Koskenniemi & Hälinen 1970). It may be group work, problem solving, or programmed teaching, and it may be either direct or indirect. In the past, in the practice teaching of physical education, attention has been directed mainly to the contents of programs, while the development of forms of teaching has occupied a secondary position.

1.2 Teacher education research

Dunkin (1987) refers to a key statement by Gage (1972) "Teacher education is one context in which teaching occurs. It is an especially interesting context because teaching is the basis of objectives guiding teacher education programmes, as well as a process by which those objectives are attained and main outcome by which the success of programmes is judged" (8).

The pedagogical and didactic problems of teacher education are a special subarea of what is now frequently referred to as the "pedagogy of higher education". The Finnish National Commission on Teacher education (Vuoden 1973 opettajankoulutustoimikunnan mietintö, 1975) suggested that the most important sectors of research and the pedagogy of higher education concern (a) the problems of the overall aims of higher education, (b) the problems related to the development and investigation of instruction, and (c) the special problems of educational technology and teaching methods. Within this latter area of concern, teacher education, one of the key issues is practice teaching. Researchers and teacher educators are constantly faced with the problems of

how the teaching practice experience should be planned and developed so that the intended competences can be optimally attained.

In January 1974 the Department of Physical Education at University of Jyväskylä introduced, on an experimental basis, a new type of practice teaching using microteaching based on the 'Human Interaction Model' (Flanders 1987, 20). The new preservice course that emerged formed part of the degree requirements and was intended to be given during the last term of the third year as an obligatory course (45 hrs). At the same time also a course of didactic observation was introduced to the curriculum of the faculty (30 hrs) given before the course of microteaching. After the study reform (1978-) these courses were combined (2 study weeks) and given in the last term of the second year and the first term of the third year. Asetus liikuntatieteellisestä tutkinnosta no 299 [Examination requirements of Department of Physical Education University of Jyväskylä, no 299]. Jyväskylän yliopiston opinto-opas [Study quide of the University of Jyväskylä] (1975-, 1979-) (see Telama 1975, Telama & Vuolle 1976, Telama et al. 1980).

The planning and implementation of the course necessitated a meta-level framework on the concepts and methods of interaction analysis. The construction of the theory-based interaction model and the related observation instrument for Physical Education, PEIAC/LH-75, (see Chapter 5), and its modification PEIAC/LH II (see Section II. Chapter 4), that were used as feedback instrument in microteaching. The final instrument was the result of empirical pilot studies (Heinilä 1970, 1971, 1974, 1977a, 1977b, 1990), based on the pioneering work of Flanders (1965, 1970) (Heinilä 1977b). Curriculum evaluation and validation of the basic elements of the course package was done during the period 1976-1990 drawing on the expertise of the Helsinki DPA project (e.g., Komulainen 1968, 1970, 1971a, 1971b, 1973, 1974, 1978, Komulainen & Kansanen 1981 (eds.), Koskenniemi & Komulainen 1969, Koskenniemi 1981). Finally, the program evaluation, an ex-post facto long-term longitudinal inquiry, was done during the period 1974-1992. (Heinilä 1974, 1976, 1977a, 1977b, 1988, 1990, 1992a, 1995, 1992b)

An adapted version of Gage's (1972) model of research illustrates the place that the present research occupies in this field (Figure 2).

As can be seen from the model, the task of the project is to identify detailed, observable teacher behaviors that are related to student learning. The task of teacher education is to help student teachers get to know, understand and adopt effective teacher behaviors. In connection with the application of scientific knowledge to teaching practice, the dimension of "Knowledge That vs. How" was the central concern in the 1970's (Gage 1978). So-called performance-based teacher education programs have been based on this outlook and the best known of such programs are didactic observation and microteaching. In such courses methods of interaction analysis have been used as a tool to help bring about changes of behavior. In connection with these methods, observation is also seen as a teaching skill. It is through the use of these methods that this study will examine both the problem of describing the nature of observation teaching and the development of techniques to study these activities and behaviors. At the same time, it will be shown that the methods of interaction

analysis provide a new basis for the selection of the forms and contents of teacher training so that the occupational demands of the teaching profession are fulfilled, and theory and practice can be brought closer together, aimed in the study reform in 1978 at the department of physical education University of Jyväskylä.

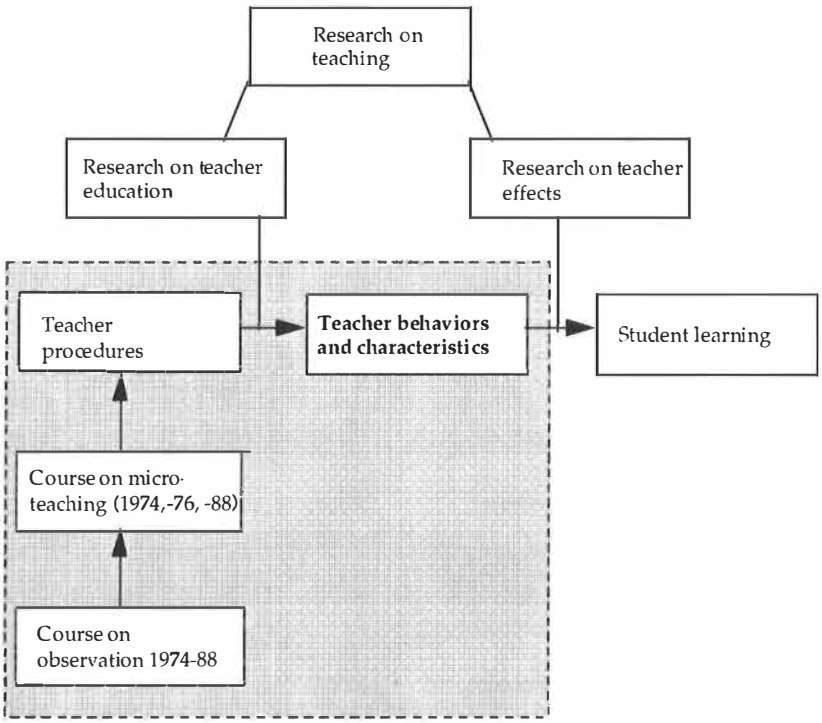


FIGURE 2 Adapted version of Gage's (1972) model of the field of research on teaching (Heinilä 1977b)

1.3 Framework of study program

As can be seen in the preface of the dissertation, this is a cumulative monograph, a summary of large number of research projects carried out in a 30-year time period. These research projects are all related to Dunkin's model (Dunkin 1987, XV), which was presented on page 25. In summary, then, the purpose of this dissertation is to report on the main findings of a research program on the use of interaction analysis in physical education. The framework of the research strategy used in this longitudinal inquiry is schematically represented in Figure 3.

As can be seen from the figure, this study consists of two main sections, and the first is realized at the meta-level, whereas the second is realized at the substantive level, as a longitudinal long-term ex-post facto empirical inquiry

with the main purpose of explanation and predictive validation of a special preservice teacher education program in contextual variation.

Thus, drawing on earlier reports, in the first section, a) the theoretical framework of the project and its relation to other work on interaction analysis will be described, b) an account of the construction of the observation instrument will be given, c) the empirical structure of the instrument will be explored, d) the measurement properties (reliability, objectivity of coding, variability of coding, and construct validity and sensitivity) will be investigated. In the second section, e) the application of the instrument in a micro-teaching program, and curriculum evaluation will be described, f) the validation of the basic elements of the revised program will be described, g) the predictive validation of the program in an longterm multi-dimensional inquiry will be presented, and finally, h) the implications of the study for further research will be discussed.

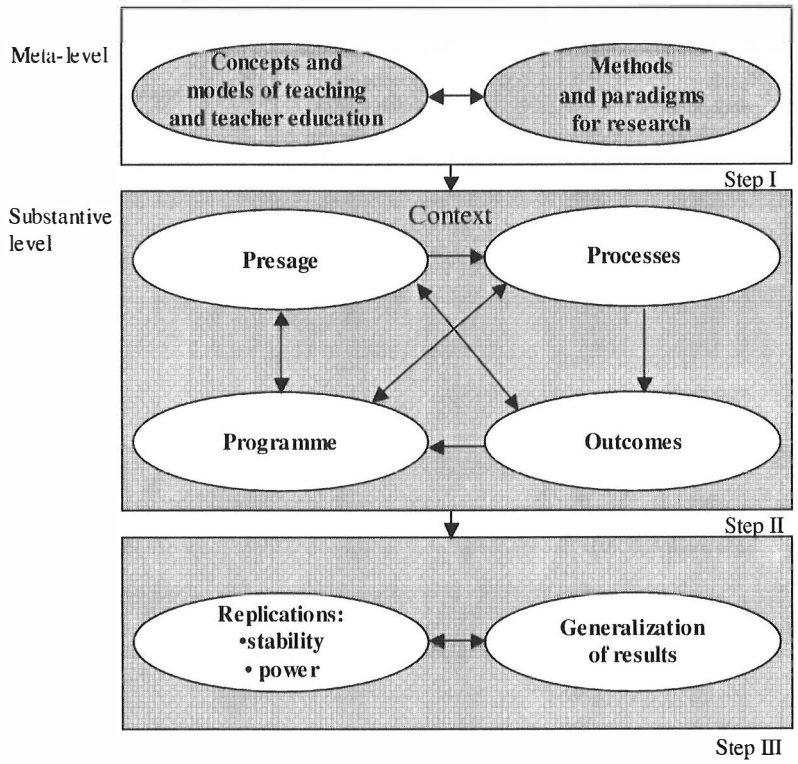


FIGURE 3 Framework: components in relation to other components and research strategy (Heinilä 1992a)

2 REVIEW OF LITERATURE

2.1 Overview

In order to set the present study in its proper context, this chapter will present a review of literature related to research on classroom observation. While Binet's dictum, quoted in the introduction, still is not much of an exaggeration as a summary of the state of education as science, it is true that some researchers in education became interested in analyzing classroom interaction as early as the late 1930's. Since that time, a number of category systems for analyzing primarily verbal interaction in the classroom have been constructed. A survey in the mid-sixties by Amidon and Simon (1965) reported twenty such category systems. Once developed, such category systems have been put to use in a great number of research studies. Early work involving systematic observation in classrooms was reviewed in the first and second editions of the *Handbook* of *Research on Teaching* by Medley and Mitzel (1963) and by Rosenshine and Furst (1973). Medley (1982) wrote a review of systematic classroom observation in the fifth edition of the *Encyclopedia of Educational Research.* And in *the International Encyclopedia of Research on Teaching and Teacher Education* (Dunkin 1987 (ed.)) Medley, Flanders and Dunkin wrote articles where the application of Interaction analysis for teacher education research was discussed from the perspective of criteria for evaluating teaching. For the International Encyclopedia of Research on Teaching and Teacher Education, Dunkin (1987) used a two level - six blocks conceptual framework represented in Figure 4.

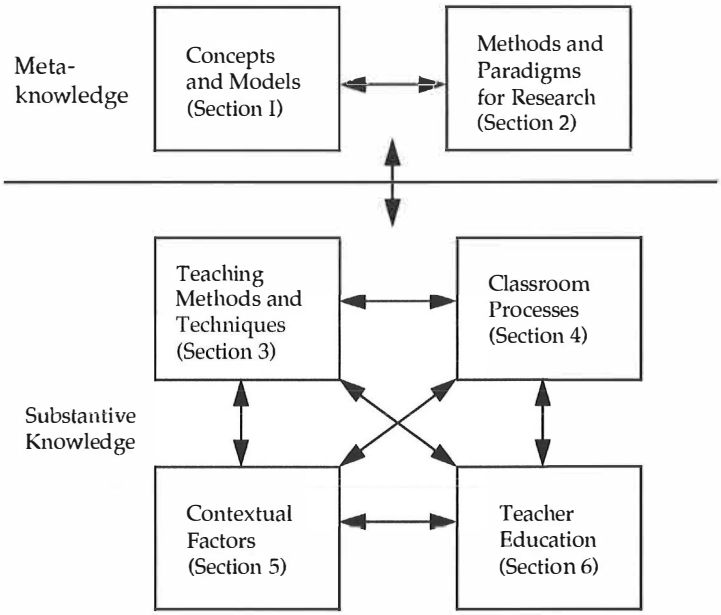


FIGURE 4 Schematic representation of the sections in relation to other sections of the international encyclopedia of teaching and teacher education (Dunkin 1987, XV)

Thus, each block of content is portrayed as interacting with the others, with "Teaching Methods" and "Techniques" and "Classroom Processes" occupying the central panel. Dunkin (1987) noted that "In this framework *teaching processes* are seen to be the subject of concepts and theoretical models of teaching and research on teaching. As mentioned, teaching processes are also seen to be influenced by and to influence their contexts and the process of teacher education." (XV). Thus, any research study of teaching and teacher education, which focuses on classroom processes, occurs within the context of a well-established research tradition.

In this chapter we will first discuss the historical development of research on teaching, including the development of interaction analysis. Secondly, the most commonly used observation system in educational research, the Flanders Interaction Analysis Category System, will be described and discussed. This discussion will be followed by a review of research in physical education, which has used interaction analysis and observation methods. Finally, these studies will be critically discussed in terms of their success in achieving valid and reliable results.

2.2 Historical development of research on teaching

In their article on observation research, Evertson and Green (1986) identified four overlapping phases of the history of this approach to the study of educational processes. Phase One (ca. 1939-1963) was an exploratory phase, which attempted to identify teacher-student interactions and other related classroom and instructional behaviors. Phase Two (ca. 1958-1973) was a period of instrument development, and of descriptive, experimental, and training studies. The use of category systems and issues about paradigms for the study of teaching emerged during this phase. During Phase Three (ca. 1973 to the time of the review) studies explored teacher behaviors that relate to student achievement, usually on standardized tests. Phase Four (ca. 1972 to the time of the review) ran concurrently with Phase Three and was a period of expansion, alternative approaches, theoretical and methodological advances, and of convergence across research directions in the use of observational techniques.

This historical review of research on teaching will attempt to explore some of the work done during these phases of study with particular emphasis on the period of expanding theoretical and methodological advances in the use of observational techniques.

2.3 Early research on teacher effectiveness

Although research on teaching, as defined by Gage (1963b, 1972), is relatively new, research on "teacher effectiveness" has been conducted for many years. The early studies were stimulated by the desire to provide an objective basis for the selection, training, employment, and promotion of teachers, but in reality they offered minimal opportunity for a real understanding of teacher effectiveness. In general, as Dunkin and Biddle (1974) emphasized, such studies revealed no more for teachers and educators than the discovery that performance on college examinations and in practice teaching were apparently unrelated to subsequent success in teaching. Many reasons have been offered by reviewers and critics for the failure of these early studies. Dunkin and Biddle summarized these as (1) the failure to observe teaching activities, (2) theoretical impoverishment, (3) the use of inadequate criteria of effectiveness, and (4) the lack of concern for contextual effects.

With the development of the behavioral sciences in the first half of this century, attempts were made to apply these scientific methods to the problems of teacher behavior, school learning and teacher education. As Dunkin and Biddle pointed out, perhaps the most significant shortcoming of these early studies was that they consistently avoided looking at the actual process of teaching in the classroom. They further suggested that if teachers vary in their effectiveness, it must be because they vary in the behaviors they exhibit in the

classroom. For this reason, the focus of a study on teacher effectiveness ought to be on the classroom where the teaching actually takes place.

2.3.1 Development of analytical research methods

During the 1960's, descriptive analytical research in general education increased considerably and became an independent branch of intellectual inquiry. Its general theoretical orientation became clearer and acquired a more definite direction. Research in this area has been directed towards (1) natural teaching situations; (2) the whole of the teacher-pupil interaction process; and (3) the construction of a uniform theoretical basis and conceptual scheme, within which the newly acquired empirical data can be placed, analyzed and generalized. (See e.g., Birkin 1971, Dunkin & Biddle 1974, Heinilä 1974, 1977b, Westbury & Bellack 1971).

This orientation was greatly influenced by the development of quantitative methods, and observation research has occupied a key position. In this context, observation research refers to the analytical methods based on observation, during which behavior is observed and classified. With this method, a classification system can be based on (1) theory, (2) a theoretical model, (3) existing observational systems, or (4) the results of empirical studies or pilot studies. When the focus of research shifted from teaching efficiency research towards the investigation of the classroom atmosphere and the regularities of the teaching-learning process, observation became the most practicable method.

2.3.2 Development of observation recording instruments

In the field of observation research, the problems of content and method are closely related and they should therefore be examined simultaneously. The use of a measuring instrument implies a theoretical base. Such is, of course, also the case with, for example, the classic interaction analysis systems by Flanders (1965, 1970) and Bales (1950). When a researcher adopts an instrument of this kind, he has not only made a methodological decision, but he has also committed himself to a particular theory and group of variables. In the study of teacher behavior, the theoretical base might be the observed variable values placed on given dimensions, such as teacher-centered/pupil-centered, direct/indirect, etc., or the description of event sequences, for instance, by means of time-line display (cf. Flanders 1970).

Analytical methods based on observation generally include (1) a group of carefully specified categories for the classification of the behavior under observation, (2) a group of standardized procedures which define the observation procedure, (3) instructions for processing, analyzing and presenting the data in a meaningful way which corresponds as closely as possible to the original events (Flanders 1970, Heinilä 1970, 1974, 1976). The category system employed will determine the number and quality of the events, which, defined in terms of interaction analysis systems, are exhaustive and mutually exclusive.

During the 1960s and 1970s, a great number of recording instruments were developed for the study of teaching. (For reviews of some of these see, e.g. Biddle 1967, Dunkin & Biddle 1974, Medley & Mitzel 1963, Rosenshine 1971, Rosenshine & Furst 1973, and Simon & Boyer 1970.) Although these instruments have a common purpose to systematically record teacher-student behavior in the classroom, there are some major differences among them. These differences relate primarily to the dimension or dimensions of the classroom activity to be recorded. Generally, the focus of the instrument reflects the theoretical orientation of the investigator. The particular orientation of the investigator not only guides the general direction of the research work, but is also the key in making decisions concerning the logical steps in the development of the system.

Simon and Boyer (1970) reported altogether 92 different recording systems, of which 73 were designed for observing classroom behavior. They suggested foci for categories within recording instruments and classified them as follows:

1. Affective - the emotional content of communication;
2. Cognitive - the intellectual content of communication;
3. Psychomotor - the non-verbal behaviors, posture, body position, facial expressions, and gestures;
4. Activity - what is being done that relates a person to someone or something else (for example, reading or hitting a ball);
5. Content - what is being talked about;
6. Sociological structure - the sociology of the interactive setting, including who is talking to whom and in what roles; and
7. Physical environment - descriptions of the physical space in which the observation is taking place, including the materials and equipment being used.

In a review of almost 500 studies involving the systematic observation of classroom teaching, Dunkin and Biddle (1974) identified six classifications according to content and/or the theoretical "orientation" toward teaching. These classifications are:

1. studies dealing with classroom climate;
2. studies dealing with management and control of pupil behavior in the classroom;
3. studies dealing with the classroom as a social system;
4. studies dealing with the knowledge and intellectual aspects of teaching;
5. studies dealing with logic and linguistics; and
6. studies dealing with the sequential patterns of classroom behavior.

Rosenshine (1971) classified the observation instruments used in fifty-one studies into "category systems" and "rating systems". In a category system, each behavior of the teacher or student was coded whenever it occurred. In a rating, or "sign" system, outsiders or students estimated the behavior of the teacher and a five- or seven-point scale. These observation systems were also classified according to the amount of inference required of the observer or the person reading the research report. The term inference refers, in this context, to the

process intervening between the objective behavior seen or heard and the coding of this behavior on an observational instrument. Category systems are classified as "low-inference" measures because the items focus on specific, denotable, relatively objective behaviors, such as "teacher repeats student's idea" or "teacher asks evaluative questions", and also because the behaviors are recorded as frequency counts. The rating systems are referred to as "high-inference" measures because they lack the specificity of low-inference variables. In general, the category systems of observation have been used most frequently. They appear to be more flexible than sign observation and rating systems, provide more data, and have a higher level of objectivity in coding (Dunkin & Biddle 1974, Rosenshine 1971). This has been well documented in the 1980s also in reviews connected to Research of Sport Pedagogy (Piéron & Cheffers 1988).

To summarize, the preceding review has indicated that a large number of observational recording instruments have been developed to investigate classroom interaction. These can be divided into "category systems" or "rating systems". The former are regarded as "low-inference" systems because of their high degree of specificity, whereas the latter are regarded as "high-inference" systems, because they operate with more general concepts.

The work of researchers involved in classroom interaction analysis was primarily motivated by a desire to prove that certain preferred interaction patterns are superior for classroom learning. The concepts "integrative/ dominative", "democratic/authoritarian", "student-centered/teacher-centered" and "indirect/ direct", all spring from a conviction that most teachers could be more effective if they would interact with pupils rather than direct them.

2.4 Development of interaction analysis

In this section, an attempt will be made to outline the basic assumptions of the traditional interaction analysis paradigm. Given this frame of reference, it should be easier for the author to present a survey of related literature in a succinct form and for the reader to follow the exposition.

Kuhn (1962) introduced the term "paradigm" to denote the fact that same accepted examples of actual scientific practice, including law, theory, application and instrumentation, all together provide models which give rise to coherent traditions of scientific research. Sharing a paradigm means that there is a shared commitment to the same rules and standards for scientific practice. Kuhn suggests that scientists work from models acquired through education and through exposure to a common core of literature. This happens often without an explicit knowledge of why the models have obtained their status. It is even possible that there is no clear-cut underlying body of rules and assumptions for the research traditions.

Kuhn's point is relevant for the interaction analysis paradigm as well. A student of interaction analysis has no single article or theoretical exposition to consult but, instead, needs to get acquainted with a number of paradigmatic articles and research studies. It is partly through such "finger exercise", as Kuhn

refers to it, that researchers learn how to implement an empirical study of classroom processes.

2.4.1 Assumptions of the traditional interaction analysis paradigm

Some key assumptions of the traditional interaction analysis paradigm are listed below.

1. A basic assumption within the interaction analysis paradigm is that the social-emotional climate influences behavior. In a school and class setting, this means that a positive social-emotional climate is beneficial for almost any aspect of education. Various researchers have used somewhat different terminology to express roughly the same basic assumption.
2. It is generally assumed that the social-emotional climate is a group phenomenon and that the teacher's behavior is the most important single factor in creating climate in the classroom.
3. The teacher's verbal behavior is assumed to be a representative sample of his total classroom behavior. As a result of this assumption, it is commonly considered sufficient to observe and record only the verbal behavior of the teacher and students in the classroom.
4. The decision to focus exclusively, or mainly, on the recording of overt verbal interaction is enhanced by the assumption that verbal behavior can be observed with greater reliability than nonverbal behavior.
5. It is assumed that the study of classroom interaction cannot be done by means of self-reports by the teacher and the students, e.g., through questionnaires or checklists. Interaction must be observed and recorded by an observer who is not simultaneously engaged in that interaction.
6. It has been assumed that observers could be trained to give a faithful record of what actually transpires in the classroom. In addition, it has been assumed that someone trained in the observation method could also decode an observation protocol and, as it were, reconstruct the interaction.

We have already discussed in general terms the development of observation recording instruments, how they have been classified and how the special terms associated with them have been defined. At this point, we will discuss some early studies based on the traditional interaction analysis paradigm. Then, since the principal indebtedness of the present study is to the Flanders system, we will describe the Flanders' Interaction Analysis Categories (FIAC) system. We will then narrow the focus to give an account of the interaction analysis paradigm within physical education. Finally we will discuss studies in physical education that have used adapted versions of the FIAC system.

2.4.2 Early studies of teacher behavior

The formal study of teacher behavior had its origin in the Progressive Education Movement, inspired by the philosopher John Dewey, under the

influence of Harold Anderson (1939) and the research group consisting of Kurt Lewin, Ronald Lippitt and Ralph White (1939). These early researchers felt a need to make classrooms more student-centered, to abandon the autocracy of education, and to promote the ideals of democracy and group dynamics. The climate of the classroom became very important.

Using the notion of a "social emotional climate", Anderson conducted systematic studies into the effects of teacher behavior upon pupil behavior. The psychological assumptions of these studies are that the child learns less if he or she is given the answers to his schoolwork, and that he grows less in other respects if the teacher makes all the decisions concerning content and procedure. Anderson quantified behavior phenomena and thus provided the basis upon which Flanders later demonstrated that indirect teacher behavior had a positive correlation with child achievement.

Dominative and integrative behavior of the teacher was observed and identified by Anderson with a category system containing nineteen categories: eleven domination categories and eight integration categories. Anderson also showed that it was possible to compute an index, or ID-ratio, by dividing the number of integrative contacts by the number of dominative contacts, and that teachers could be compared using this index criterion.

Lippitt and White (1943), together with Lewin, conducted a series, of laboratory experiments for determining the effects of adult teachers' influence and the organized and voluntary activities of boys clubs. Each club was subjected sequentially to an adult playing the role of an "autocratic leader", a "democratic leader", and a "laissez-faire leader". The results of these studies confirmed or extended the general conclusions of Anderson. As a result of these two basic and independent studies, which produced mutually supportive results, the notion of a social emotional climate was established.

Drawing upon the work of both groups, Withall, through extensive analysis, produced an index of teaching behavior, which, though almost identical with the integrative/dominative ID-ratio of Anderson, offered a much more refined category system of classroom climate. Withall (1949) defined the concept "social emotional climate" as the "general emotional factor, which appears to be present in interactions occurring between individuals in face to face groups" (p. 348). In practice, this "climate" was considered to influence: "(1) the inner private world of each individual; (2) 'the esprit de corps' of a group; (3) the sense of meaningfulness of group and individual goals and activities; (4) the objectivity with which a problem is attacked; and (5) the kind and extent of interpersonal interaction in a group" (pp. 348-349).

Withall emphasized the importance of the teacher's verbal behavior in determining the classroom climate and identified the preliminary categories of his research instrument by recording regular class sessions and analyzing tape-recorded lessons. From this analysis he devised a system of classifying the teacher's verbalization into the following seven categories:

1. learner-supportive statements;
2. acceptant and clarifying statements;
3. problem-structuring statements or questions;
4. neutral statements;

5. directive or hortative statements with intent;
6. reproving or deprecating remarks;
7. teacher self-supporting remarks. (Withall 1949, p. 349)

These seven categories ranged from 'learner-supportive' statements (1-3) through 'neutral' statements (4) to teachers' self-supporting statements (5-7).

Extensive validational procedures followed the development of this category system to determine the objectivity, reliability and validity of the climate index.

The objectivity of Withall's instrument was reported in terms of inter-judge agreement. Data for computing the indices were obtained coding teachers' statements contained in three typescripts, and the percentage of agreement of each of four observers with the investigator was computed. The percentage agreement of each observer with the mean percentage of agreement ranged from 56% to 75%.

Reliability was evaluated by determining the consistency of the instrument. Day-to-day variations in the pattern of statements of three teachers were compared. The Chi Square test was used to check the hypothesis that no significant differences occurred from day to day.

To determine the validity of the climate index, four procedures were used: (1) Anderson's Teacher Behavior Categories as the criterion instrument; (2) pupil evaluations; (3) a Teacher Characteristics Rating Scale; and (4) the description of the class situation from three frames of reference.

As a result of these studies, and later those of Ned Flanders (1965, 1970), the school of interaction analysis was created (Amidon & Hough 1967).

2.4.3 The Flanders interaction analysis category system (FIAC)

Clearly, the research instrument most often used in classroom studies is the Flanders Interaction Analysis Category System (FIAC) and its modifications (Dunkin & Biddle 1974). This system is based a social psychology and the theory of the leader/subordinate relationship. A knowledge of Flanders' studies and of interaction analysis is important to the understanding of this particular approach to evaluating measuring instruments, since the choice of a system of classification, as well as decisions concerning its modification, involves adherence to a theoretical frame of reference as its basis.

According to Flanders (1970), the main goals guiding the analysis of teaching behavior are (1) to help the teacher develop and control his teaching behavior, and (2) to investigate relationships between classroom interaction and teaching acts so as to explain some of the variability in the chain of events. Flanders defined an event in terms of time: whatever goes on during a three-second interval is treated as one event and coded as such. With this in mind, Flanders' theory is an attempt to explain teacher influence and changes in pupil behavior, in which the intervening hypothetical mechanism is the process of goal clarification. Accordingly, teaching is a process of clarifying and implementing objectives, in which the teacher's task is to act flexibly so that there develops a minimum of dependence in pupils (Flanders 1967b). In developing his theory, Flanders introduced some basic changes to classroom

research by reconceptualizing the continuum of teacher behavior variability, by moderating Anderson's (1939) "Commitment" in which classroom democracy was always advocated and domination avoided, and by including in his new observational instrument additional categories for judging pupil verbal behaviour.

Definition of terms

The following concepts were used in describing tentative hypotheses of teacher influence (Flanders 1967b).

Direct influence consists of stating the teacher's own opinion or ideas, directing the pupil's action, criticizing his behaviour, or justifying the teacher's authority or use of that authority.

Indirect influence consists of soliciting the opinions or ideas of the pupils, applying or enlarging on those opinions or ideas, praising or encouraging the participation of pupils, or clarifying and accepting their feelings.

The word *dependence* refers to the essential qualities of a superior-subordinate relationship. The opposite of dependence is independence. *Independence* refers to a condition in which the pupils perceive their activities to be "self-directed" (even though the teacher may have helped create the perception) and they do not expect directions from the teacher. It is assumed that various degrees of dependence or independence exist.

High dependence refers to a condition in which pupils voluntarily seek additional ways of complying with the authority of the teacher.

Medium dependence refers to the average classroom condition in which teacher direction is essential to initiate and guide activities but the pupils do not voluntarily solicit it. When it occurs, they comply.

Low dependence refers to a condition in which pupils react to teacher directions if they occur, but their present activities, usually teacher initiated, can be carried on without continued teacher direction. In the face of difficulties, pupils prefer the teacher's help.

Dimensions of classroom learning and teaching

One aspect of the classroom situation that should make a difference in the pupil's reaction to teacher influence is his perception of the learning goal and the methods of reaching that goal. One can conceive of a situation in which the goal and the methods of reaching the goal are clear to the pupil, and another situation in which these are unclear. Certainly, when a student knows what he is doing, his reactions to teacher influence will not be the same as when he is not sure of what he is doing. The student may also perceive the goal as desirable or undesirable. The attraction of a goal determines motivation, an attribute, which Lewin (1935) designated as positive valence or negative valence.

Changing the mode of teacher influence (direct-indirect) along with the process of goal clarification Flanders (1967b) calls "flexibility". Flexibility of teacher acts may explain why direct influence may increase or maintain

dependence in one situation, and increase or maintain independence in another. Flanders presented tentative hypotheses of teacher influence, which can be illustrated in the following way (Komulainen 1973):

Mode of teacher influence	Process of goal clarification		
	Unclear	Clear +	Clear -
Direct	+	±	+
Indirect	-	-	-

+ = dependence increases
- = dependence does not increase
(Flanders 1967b, 110-116)

In a different context, Soar (1968) showed that the level of difficulty of the subject matter presupposes that the teacher uses different modes of influence or flexibility. Creative activity demands a freer setting and less control in order to be optimally successful. Thus, the structure of the subject matter is an important factor in determining authority in use.

As Flanders (1987) pointed out, twenty years later in process-product experiments, the correlation between "x" and "y" was in most cases nonlinear. This means that there is some intermediate level of "x" which produces a maximum "y", but too little or too much "x" will reduce "y" (p. 22).

Later, Flanders (1970) added to his theory the domain of social access, which consists of social contacts and the range of ideas. The presumption of social access for communication means that most of what takes place in the classroom depends on communication. Who talks to who forms a network of communication, which is closely related to physical access, such as the seating arrangements in a classroom. The opportunities to contact other pupils can be at a minimum when the formation is restricted, whereas if mobility permits pupils to select their communication contacts the formation is free. When the ideas discussed are determined primarily by the teacher, the range of ideas is controlled, and when anything can be discussed, the range of ideas is open. "In most instances, free social contacts also permit a wide range of ideas to be discussed" (Flanders 1970, p. 315).

The measurement of social contacts can be made by asking observers to make a separate assessment of class formation and to record notes whenever this formation changes. Similar evaluation of the range of ideas can be made by using pupil questionnaires to determine whether the pupils' perceptions about expressing ideas is controlled or open.

The possible configurations of these four dimensions of classroom teaching and learning, i.e., goal orientation, authority in use, social contacts, and range of ideas, is illustrated by the use of the following Figure 5.

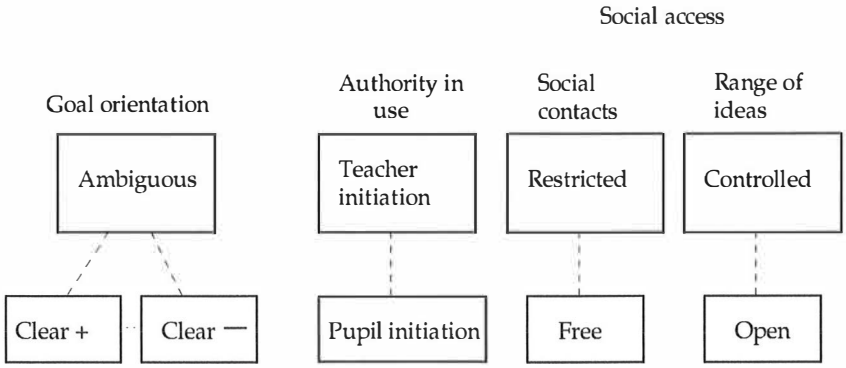


FIGURE 5 Flanders descriptive model (Flanders 1970, 317)

Knowing the sequence and variety of the possible configurations in the four domains discussed can help to predict what will happen next. Flanders used the term *variety* to refer to the total number of different configurations, which may occur in the classroom and the term *sequence* to indicate how many different configuration pairs occurred in a given period.

Flanders summed up his hypotheses concerning the conditional relationships, which predict educational outcomes in the following manner:

If... a certain goal orientation exists
(here we begin with the pupil's goal perceptions)
And... classroom interaction is characterized by
a) certain authority in use
b) certain social contacts
c) and range of ideas social access
(here are features of the interaction)
Then we probably
Expect certain educational outcomes, in terms of
a) pupil initiation and self-direction
b) average pupil attitudes
c) average subject matter achievement
(Flanders 1970, 320)

The Flanders observation instrument

On the basis of his theories, Flanders developed a new observation instrument which was in some ways an improvement to earlier ones and more useful, e.g., as a means of teacher training. Referring to the classifications of Simon and Boyer (1970) mentioned earlier, the Flanders Interaction Analysis Category System focuses upon the first classification, "affective". But, as Flanders points out, it emphasizes both the affective and the cognitive domains in the classroom. In spite of his emphasis on the classroom climate, Flanders was very much aware of the role of the cognitive domain in the classroom. "Every pattern of interaction has a cognitive and an affective component. To understand what

goes on in the classroom is to take both into consideration" (Flanders 1970, 270 and 1987, 20).

Building on Withall's learner-centered/teacher-centered continuum, Flanders identified his teacher talk categories as representing indirect/direct behaviors. Categories 1, 2, 3 and 4 were considered indirect behaviors and categories 5, 6 and 7 represented direct behaviors (Table 1). The continuance of the indirect (integrative)/direct (dominative) dichotomy, introduced by Anderson earlier, also allowed Flanders to compare teachers in terms of ID-ratios.

The analysis of "initiative" and "response", a characteristic of interaction between two or more individuals, is the major feature of Flanders category system (Table 1). "To initiate", in this context, means to make the first move, to lead, to begin, to introduce an idea or concept the first time, to express one's own will. "To respond" means to take action after an initiation, to counter, to amplify or react to ideas which have already been expressed, to conform or even to comply to the will expressed by others. Flanders (1970) suggested that the teacher is expected, in most situations, to show more initiative than the pupils. His category system was intended to be used to study the balance between initiation and response. He pointed out that a different category system would be needed to investigate other problems of teaching and learning, such as, the effect on class learning of different pupil reactions.

With seven categories of teacher talk and only two of pupil talk in the FIAC system, more information is provided about teachers in general, and therefore how teacher statements influence the balance of initiative and response behavior can be studied only with a particular set of these categories. In general, the quality of the statements is associated with educational outcomes just as much as, if not more than, quantity.

By using Flanders' system, it is possible to identify the quantity and relationship of pupil talk and teacher talk, to classify teacher-pupil behavior, and to record a sequence of verbal events in live classroom situations. The sequence of verbal events can then be displayed in matrix form where frequencies and relationships of various teacher and pupil verbal behavior patterns may be ascertained. Following Darwin, Flanders has also considered matrices as first order Markov Chains in order to compare two matrices (Darwin 1959, Flanders 1967a). Similar methods of observation and analysis of data have also been applied by Bales (1950) and Pankratz (1967).

TABLE 1 Flanders interaction analysis categories (Flanders 1970, p. 34)

	Flanders Interaction Analysis Categories* (FIAC)	
Teacher Talk	Response	1. *Accepts feeling.* Accepts and clarifies an attitude or the feeling tone of a pupil in a nonthreatening manner. Feelings may be positive or negative. Predicting and recalling feelings are included. 2. *Praises or encourages.* Praises or encourages pupil action or behavior. Jokes that release tension, but not at the expense of another individual; nodding head, or saying "Um hm?" or "go on" are included. 3. *Accepts or uses ideas of pupils.* Clarifying, building, or developing ideas suggested by a pupil. Teacher extensions of pupil ideas are included but as the teacher brings more of his own ideas into play, shift to category five.
		4. *Asks questions.* Asking a question about content or procedure, based on teacher ideas, with the intent that a pupil will answer.
	Initiation	5. *Lecturing.* Giving facts or opinions about content or procedures; expressing *his own* ideas, giving *his own* explanation, or citing an authority other than a pupil. 6. *Giving directions.* Directions, commands, or orders to which a pupil is expected to comply. 7. *Criticizing or justifying authority.* Statements intended to change pupil behavior from nonacceptable to acceptable pattern; bawling someone out; stating why the teacher is doing what he is doing; extreme self-reference.
Pupil Talk	Response	8. *Pupil-talk-response.* Talk by pupils in response to teacher. Teacher initiates the contact or solicits pupil statement or structures the situation. Freedom to express own ideas is limited.
	Initiation	9. *Pupil-talk-initiation.* Talk by pupils which they initiate. Expressing own ideas; initiating a new topic; freedom to develop opinions and a line of thought, like asking thoughtful questions; going beyond the existing structure.
Silence		10. *Silence or confusion.* Pauses, short periods of silence and periods of confusion in which communication cannot be understood by the observer.

* There is *no* scale implied by these numbers. Each number is classificatory; it designates a particular kind of communication event. To write these numbers down during observation is to enumerate, not to judge a position on a scale.
(Flanders 1970, 34)

To be concerned with interaction is, as Flanders (1987) pointed out later "to focus on the continuous stream of behavior, which occoures in the classroom as a series of individual acts. An act might consist of a teacher contacting the class or single student, a student contacting a teacher, a student contacting another student or student acting upon an object (after Piaget) in order to formulate knowledge" (p. 20). In this comment concerning the interaction process Flanders points the meaning of the social form of classes.

Flanders has summarized his own seven research projects on social emotional climate together with sixteen other projects that had used his 10-category observation system as a base for investigating pupil learning or behavior with an interaction analysis variable. The results obtained by Flanders tended to support the existence of a consistent, causal and often-significant relationship between teacher behavior, as quantified by the FIAC system, and the social emotional climate, as measured by attitude scales. Both of these in turn appeared to relate to achievement.

The percent of teacher statements that make use of ideas and opinions previously expressed by pupils was directly related to average class scores on attitude scales of teacher attractiveness, liking the class, etc., as well as to average achievement scores adjusted for initial ability. (Flanders 1970, 424, Flanders & Simon 1970, 1426)

In order to assess the effects of classroom interaction, Flanders (1970, 354-356) referred to the reports of 18 research projects, the purpose of which had been to investigate at different levels of education the effectiveness of using interaction analysis as a means to facilitate learning. A general objective of such programs had been an awareness of teaching behavior and the development of flexible teaching behavior. The findings of these research projects gave rise to the following generalizations:

1. An individual becomes more responsive to pupil ideas ... by learning how to code with categories of interaction analysis and by interpreting displays from specimens of his own teaching and the teaching of another person.
2. Teaching behavior becomes more flexible (or variable) as a result of studying interaction analysis.
3. The attitudes of college students toward teaching and programs of the preparation of teachers become more positive for those who study interaction analysis compared with those who don't. (Flanders 1970, 354-356)

And finally, Flanders (1970) presented the following two important comments concerning comparison between research projects in this area.

> "First, a concerted effort to continue this trend should be maintained, because this is likely to bring educational researchers closer to classroom problems and may have lasting effects on the pre-service and in-service education of teachers. Second, replications of all research projects should be promoted, not discouraged, either by graduate degree requirements as by funding organizations. It is only when a particular finding has been replicated several times with first-class research skill that we can increase our confidence in its validity." (p. 427)

Summarizing sport science studies in "Research in Sport Pedagogy from empirical analytical perspective" Piéron and Cheffers (1988, 197) found that the first comment concerning the concerned effort presented by Flanders (1970) had been realized to be fruitful in the area of research on sport pedagogy but that the second one still needed to be promoted.

2.5 Interaction analysis in physical education research

Although descriptive analytic research involving interaction analysis has gained considerable popularity among educators over the past few decades, physical educators for the most part failed to acknowledge the benefits of such research in 1960's and 1970's and this fact was one of the mains motives for starting this longterm research project. In more than a hundred studies reviewed by Dunkin and Biddle (1974), which had dealt with applications of the FIAC system and related instruments, none of them were used in the context of physical education.

After reviewing 700 American descriptive-analytical studies on physical education, Nixon and Locke (1973) concluded that such research was in its infancy in the early seventies and had only begun to come to grips with the problems and prospects of fruitful investigation. It had consisted mainly of fairly unsystematic surveys of various features of teacher-pupil interaction and had generally been colored by attempts to improve the effectiveness of teaching. The focus of these surveys had been sometimes on the teacher, at other times on the pupils, and again on particular behaviors of both parties, such as teacher talk, pupil movement, contents of physical education, etc.

In physical education research, there was in the 1970s a total lack of a unified theoretical basis, or even a general model of the teacher-pupil interaction process. This was considered as a serious drawback, which was seen to slow down the progress of research. As Nixon and Locke noted, "it has been difficult to classify, evaluate and co-ordinate investigations" (Nixon & Locke 1973, 1129). As a result, our knowledge of teacher-pupil interaction in physical education was rather modest in the seventies. (See, e.g. Anderson 1971, Locke 1977, Mosston 1966). But, progress during the period 1976-1984 in this area was evident (Hanke 1978, 1980b, Locke 1983, 1984). Piéron (1986) observed at the 1984 Olympic Scientific Congress in his article "Analysis of the research based on observation of the teaching physical education", that

> "In the last decade a specific area of research in teaching physical education has developed. It is based on a systematic observation of teacher behavior, student behavior, teacher-student interaction and the contextual aspects of teaching. It has led to a better understanding of the teaching act. It has helped methodologists to move beyond sole reliance on their subjective impressions and to base some of their recommendations on data derived from research rather than on unrealistic romantic expectations from programs" (Piéron 1986, 193).

This progress was reflected in programmatic research on teaching physical education in a broad international context at the Universities in the United States, Europe, Brasil, Egypt and in Canada. It has been well documented and analysed by Hanke (1979, 1980b), Darst et al. (1983), Locke and Dodds (1981), Locke (1983), Piéron (1982a, 1983a, 1984, 1989, 1993, 1994, 1996), Piéron and Cheffers (1988), Silverman (1991), Silverman and Skonie (1997). However, based on brief reviews in of research USA 1960-1980, 1980-1982 on teacher education Locke (1983) noted that "the body of knowledge and a domain for inquiry in physical education, teacher education remains uneven, unpopular and largely unread" (p. 285). He argued that "our near total failure to examine the social and psychological context of teacher education from perspective of participants is the main impediment to its improvement" (p. 285).

In the trend analysis presented by Piéron (1984), the category system used consisted of the following seven dimensions: (1) the main object of the study, (2) the subject matter taught, (3) the type of document, (4) the type of study, (5) the teaching population, (6) the student and learners, (7) context and program. E.g. in the last dimension the concepts of presage, context, process and product as defined by Dunkin and Biddle (1974) and program variables as defined by Tousignant and Brunelle (1982a) were used. This analysis based on 216 selected research reports gave a good description of the research trends until 1984: the documents analyzed in the study were distributed as follows: teacher - student interaction (30,1%), teacher behavior (28,7%), teacher behavior modification (13,9%), student behavior (11,1%), process-product (6,9%), coach behavior (5,6 %), and miscellaneous (3,7%) (p. 195). About half of these studies (47,7%) focused on observation of in-service teachers, one quarter of the research topics dealt with comparative studies of teaching process according to context variables (grade level, gender, environment, ethnic groups, class size and equipment) and only 15,8 % of these studies focused on program variables. Teaching methodology was studied more frequently than other program variables. Also, a drawback was found: these kind of studies were limited to the unique observation instrument, CAFIAS, (Cheffers 1973) and this limited the possibility to comparisons as Piéron suggested (1984, 198). 62,2 % of the studies was found to provide no information for identifying the subject matter taught and, according to Piéron: "In physical education it is hard to believe that teaching hula hoop, basketball skills, dance, creative movement, or sports, gymnastics routines lends itself to the same kind of interaction or teaching behavior" (p. 199). Based on results of this "trend analysis" Piéron (1984) concluded e.g., that further improvement of actual body of knowledge could be made in focusing on program variables rather than on less meaningful context variables. However, the doors had been opened wide to develop process-product studies in contextual variation and to link quantitative observation with qualitative appraisal. But, in the 1980s, the descriptive-correlational-experimental loop referred to by Rosenshine and Furst (1973) was far from being completed in physical education teaching and teacher education research" (Piéron 1984, 201).

2.6 Observation instruments in physical education research

2.6.1 Overview

Based on reviews referred to and trend analyses it can be stated that in the past few decades there has been also a growing interest to construct measuring instruments for the observation of the teacher-pupil interaction in physical education classes. The first attempts to construct observation instruments occurred during the 1960s (Barrett 1969, Bookhout 1967, 1969, Galloway 1962, 1970, Gorman 1969, Levin 1968). The importance of empirical descriptive research and observation instruments based on theoretical standpoints was clearly recognized in the early 1970s at a broad international level, at the universities in USA, (Cheffers 1972, Dougherty 1970, Fishman & Anderson 1971, Mancuso 1972), and in Netherlands, Amsterdam (Kemper et al. 1976), Germany, Heidelberg (Hanke 1976), Belgium (Piéron 1978b) and in Finland, in Helsinki and Jyväskylä Universities (Heinilä 1970, 1971, 1974). The results and experiences gained from the relatively few studies before 1976 were suggestive of new directions for developing the observation instrument as Locke (1977) noted in his article: "New hope for dismal science!". And based on trend analyses by Piéron (1983a) "in this area of study, a strong and growing interested was recognized at broader international level: e.g. in the year 1976 two international congresses was organized (1) in Finland, Jyväskylä, with the thems "Evaluation in development of physical education" with a section devoted to teacher-pupil interaction and evaluation of teaching process, and (2) in Canada, Québec, "Pedagogy of Sport", dealing also with teacher-pupil interaction and observation". This research trend was also well represented in the AIESEP International congress in Madrid, 1977 with the central focus on teacher education for the first time in this area. This trend has been recognized to be dominating in 1980's and 1990's (see Locke 1983, Paré 1986, 1995, Piéron 1983a, 1984, 1989, 1994, 1993, 1996). The proceedings of AIESEP (Association Internationale des Ecoles Superieures d'Education Physique) Congresses and AIESEP Yearbook have documented reports concerning these multiple research results.

During the past few decades a great number of observation instruments have been developed for research of teaching and teacher education (for reviews of some of these, see Anderson & Barette 1978, Cheffers & Mancini 1978, Darst et al. 1983, Hanke 1980b, Locke 1977, Piéron 1983a, 1994, Piéron and Cheffers 1988). Again Flanders' FIAC system and its modifications have been used most frequently (Cheffers & Mancini 1978, Piéron 1983a, Piéron and Cheffers 1988).

In the development of these instruments, perhaps the most crucial question has been to decide to what extent the original Flanders category system should be extended. How many categories, subdivisions, and/or dimensions are needed to get an adequate description of the interaction process in physical education classes and on the other hand how many are feasible? How should the adapted, extended category system be used to gain objective

coding results? These questions were answered in different ways by investigators in the 1970s, presumably because their modified observational instruments had been constructed for different purposes. It is useful to review these instruments in terms of the features, which were modified, such as content, format (number of dimensions), categories and subdivisions, as well as conceptual posture, units of analysis, and the methods used for determining the reliability and validity of the instruments.

In most cases, the purpose of the investigators in constructing these modified category systems has been to develop and test an instrument for objective observation in order (1) to describe the teaching-learning process in physical education classes (e.g. Cheffers 1973, Heinilä 1971, 1974, Nygaard 1978, Splinter 1980, Splinter et al. 1979, Tavecchio 1977), or (2) to train teachers (e.g. Cheffers 1978, Galloway 1966, 1970, Hanke 1976, Heinilä 1977b, 1990, Love & Roderick 1971, Mancuso 1972, Underwood 1979, 1980), or (3) to investigate relationships between activities in physical education classes and student growth (eg. Dougherty 1970, Gasson 1971, Kemper et al. 1976, Lamarre & Nygaard 1977, Mancuso 1972, Melograno 1971). It should be noted that all investigators have considered it necessary, as a prerequisite of validity, to extent the original single-dimensional FIAC system by adding one or more categories or subdivisions for observing the teacher's non-verbal purposeful activities as well (see also Cheffers 1973, Gasson 1971, Hanke 1976, 1980b, Heinilä 1970 and Splinter 1980) and nonverbal purposeful and non purposeful behavior (Mancuso 1972).

2.6.2 Authors using Flanders interaction analysis system FIAC or its adaptations draw patterns of teaching from their data and are describing the interaction patterns

Galloway (1962) was the first to attempt to construct an observation instrument for physical education studies. After an extensive analysis for determining the best system for the measurement of nonverbal behavior, he concluded that "no satisfactory procedure for describing nonverbal communication had until that moment been developed" (p. 7). He pursued the topic further and developed an observation instrument based on the FIAC system, which was designed to enable an observer to use the categories, time intervals and ground rules of the original Flanders system while recording the nonverbal dimension as well (Galloway 1968, 1970, 1971). The new instrument included a procedure for recording nonverbal cues associated with six of the seven teacher behaviors of the Flanders 10-category system. Double coding was used for each behavior recorded, a verbal code from the Flanders system and a nonverbal code from the Galloway system.

Dougherty (1971) used a modification of the FIAC system to discriminate between patterns of teaching. The purpose of this study was to compare the effects of Command, Task, and Individual Program styles of teaching on the development of physical fitness and learning of selected motor skills. The sub-problems were (1) to determine whether a trained observer could, using a modified FIAC system, differentiate between the three styles of teaching used

in the study, and (2) to descriptively analyze student attitudes toward the tested styles of teaching.

For the purpose of the study, an eleventh category, "meaningful nonverbal activity", was added to the Flanders system. In addition, the teacher talk categories were subdivided into interaction with the entire group and interaction with individuals. This dimension was not entered into the matrix analysis. A single trained observer was used in this study and no information was provided on the objectivity of the observer or on the validity of the revised system. However, the scores from the observations were subjected to analysis of variance. The results for the differences among the styles of teaching indicated that the Task and Individual Program groups had significantly higher ID-ratios than the Command group. It was not, however, possible to differentiate between the Task and Individual Program styles.

Gasson (1971) described the unique setting of physical education as follows:

1. the response or pupils is mainly motor as opposite to verbal,
2. the children are not static but are constantly moving,
3. there are constant changes in spatial relationships between teacher and class,
4. most primary children are eager to move and participate in concrete activities and consequently have a positive attitude toward physical education,
5. the scope of pupils' response is broader than the normal classroom with non-verbal dimension being dominant (p. 3).

For observing this setting, Gasson developed a three-dimensional observation instrument. The instrument used 22 categories to record the verbal behaviors of the teacher and pupils, the location of the teacher, and the nature and amount of child activity. To determine reliability, a "three way checking" was used. That is, the data was obtained in repeated exploratory interobserver reliability tests between himself and two trained observers, using Scott's coefficient. An interobserver reliability of .70 was reported and minimum reliability coefficients were obtained in each of the three dimensions. From the results of this study, Gasson concluded that (1) a reliable instrument had been developed, and (2) there were some indications that some teachers' verbal behavior related significantly to child activity and attitudes.

Mancuso (1972) conducted a study to determine the validity and reliability of an observation instrument, which combined the FIAC system with the Love-Roderick (1971) system. To the resulting eleven partly subdivided categories describing the teacher's verbal and nonverbal behavior, she added five categories describing pupil behavior. This single-dimensional system contained 26 categories in all. The data were gathered from simultaneous observations of three observers during a twenty-minute teaching span in a secondary physical education fencing class. A time interval of three seconds was used in coding. The reliability of the instrument was calculated by using Scott's coefficient. Reliability coefficients of .92, .91 and .92 were obtained for the three pairs of observers. (The investigator assumed the instrument to be valid

because it was based on Flanders' instrument, which was already validated. She concluded, however, that the developed instrument was in need of refinement.)

Underwood (1979) developed a single-dimensional interaction analysis system containing nine categories. The first four, Teacher Talk, Demonstration, Class Talk and Class Movement, were subscripted as "response" and "initiate". In addition, there was a category of "inactivity". Underwood used two trained observers for live situation recordings. A reliability coefficient of .96 was calculated using Scott's method on data obtained in one lesson recording.

In their studies, Nygaard (1978) and Lammare and Nygaard (1977) used the FIAC system in its basic, unaltered form. They concentrated on analyzing only verbal behavior, applying the system to the observation of audiotaped material. No information concerning reliability was supplied.

The single-dimensional category system (PEIAS) developed by Kemper et al. (1976) contained 17 categories, three of which were identified as Pupil Talk, Actions, and Performances and Demonstrations. In connection with this system, a specially developed computer program was applied for sampling videotaped behavior in real time. (Observers coded the displayed behavior by pressing a key on the keyboard of a teletype connected online with a LAB 8/e computer. The computer was programmed to record every one-second interval that the key was "on" until the observer pressed another key.)

The reliability of the instrument was determined by using Scott's *pi*. The objectivity of the instrument was operationalized as the degree of interobserver reliability and was assessed with the help of the Kendall coefficient of concordance, *W*. Three categories yielded a value of *W* significant at the .05 level, and twelve a value of *W* significant at the .01 level. Only two categories yielded a non-significant value of *W*.

The authors note that PEIAS was not standardized or validated. Therefore, it was not possible to indicate the absolute position of the teacher on the continuum directive/non-directive, and consequently, it was not possible to say anything definitive about the meaning of interteacher differences. They concluded that it was not known which ratio between directive and non-directive teacher behavior is most conducive to learning in physical education. This analysis was continued using generalizability studies (Splinter 1980, Tavecchio 1977).

Cheffers' validation study

None of the preceding studies have attempted to test the validity of their modifications of the Flanders instrument. Cheffers (1973) is a notable exception in that he has conducted a comprehensive study which concerns itself with the validation of an instrument designed to expand the FIAC system to describe nonverbal interaction, different varieties of teacher behavior, and pupil responses in physical education. In adapting the FIAC for use in physical education classes, he cited three major limitations on the original system, which prevented researchers from identifying the patterns of teacher-pupil interaction during physical education classes:

1. it is concerned only with verbal behavior;
2. it concerns itself with the classroom teacher as the sole body involved in the teaching process; and
3. without ground rule provision, FIAC describes only classes which are conducted in traditional teacher-pupil interaction on a traditional basis without regard for such class structuring as individualized learning and group activity.

The purpose of Cheffers' study was to determine whether his adaptation (CAFIAS) was valid in describing physical activity lessons with greater representativeness (content validity) than the Flanders system. Cheffers' Adaptation of the Flanders Interaction Analysis System (CAFIAS) was a 2-category system allowing the coding of behaviors as verbal, nonverbal, or both. In Cheffers' model, the teaching function was not limited to one individual (the teacher), but was identified as either the classroom teacher, another student (coded S), or the environment (coded E). To indicate group or individual teacher interaction, he simply placed either a W (whole), a P (part) or an I (not influencing) beside the relevant code symbol. A 5-second time interval was used in coding.

For a full analysis, CAFIAS required a 60x60 matrix, which Cheffers reduced to a more workable 20x20 matrix, instead of the Flanders 10x10 matrix. This comprehensive matrix was constructed to describe student behaviors as being predictable, analytical and game playing, or unpredictable and student initiated. CAFIAS was thus meant to be a very flexible research instrument for use in describing educational situations.

Six student volunteers coded the lessons for reliability testing after receiving 15 hours of training to guarantee their proficiency in the use of the new multiple category system. Three of the students used the original FIAC system, and three students used the new CAFIAS along with the investigator. The reliability was estimated by determining the interobserver agreement when lessons were coded using either of these systems. The reliability was then determined by submitting cell rankings to Kendalls' coefficient of concordance, *W*, and comparing the matrices of the student observers with those of the two main observers. Two comparisons were made, one comparing the main cell (n=10) and the other comparing the total matrices (n was specified 20x20).

All matrices developed for both FIAC and CAFIAS were reported to be concordant to the .05 level of significance and beyond. In two lessons, the badminton lesson and the creative dance lesson, the CAFIAS matrices were significant at the .05 level. All remaining matrices were significant at the .01 level. On the basis of these findings, the instrument was evaluated to be reliable.

Measures of face, content and construct validity were made possible by comparing the scores of trained interpreters answering a questionnaire (PAQ). In order to measure the performance of CAFIAS against FIAC, matrices were developed from six carefully selected physical activity classes and were presented to the interpreters. These interpreters were students who were not familiar with either system and interpreted the lessons solely from the information provided by the matrices (known as a "blind" interpretations). This

"live" interpretation group served as the control group, allowing comparisons to be drawn between their scores and the scores recorded on PAQ by the two experimental groups. It was found that the control group (outside criterion) scored significantly higher in all interpretations. CAFIAS interpreters were significantly more accurate than FIAC interpreters on the total questionnaire (PAQ), on those questions relative to CAFIAS, and on three of the films of those questions relative to both systems.

Cheffers concluded that observers are able to more accurately interpret physical activity classroom behavior when given a CAFIAS matrix than a FIAC matrix. It also appears that matrices prepared by observers working exclusively on the nonverbal dimensions were not as accurate in representing classroom behaviors as matrices prepared by observers viewing lessons in both verbal and nonverbal dimensions. Cheffers also concluded that further tests were needed to determine the sensitivity and feasibility of the instrument for use in physical activity classrooms, such as, e.g., computer programs to make multiple coding systems feasible.

2.6.3 Summary

Some observation instruments were developed in the 1970s' for use in physical education and teacher education studies. The Flanders' system has been applied most frequently and has been modified to a significant extent by varying the coverage, method of data collection and coding procedures, as well as the conceptual posture used. Modifications of Flanders FIAC system have been made by researchers in order to increase the usefulness and sensitivity of the instrument in teaching and teacher education. Modifications have been made mostly to record also nonverbal behaviors of teacher and pupils and to record different teaching patterns and teaching skills. When measuring the affective domain, the results from these instruments have been reported in terms of the basic continuum, direct-indirect influence.

Although multidimensional systems have been used most often in physical education studies no hypotheses have been presented about the relationships between clusters nor have generalizations from these relationships made. Correlative techniques were not used to analyze the relationships between the scores of categories of different clusters. The sequence and variety of teaching behavior were analyzed in only a few studies (e.g. Cheffers 1973, Dougherty 1970). Teaching behavior based on a theoretical model was discussed rarely and only in connection with verbal behavior.

In general, the investigators have considered only observer agreement and have neglected the study of validity. The validation process used by Cheffers with his multidimensional observation instrument (CAFIAS) has been discussed as an example of complicated validational procedures using different types of measurement to determine the degree of face, content and construct validity.

Reviews and trend analyses connected to sport science studies in 1980's and in 1990's have, however, shown that the knowledge concerning observation instruments and the use of them in research on teaching and teacher education

in physical education has advanced (Piéron & Cheffers 1988, Piéron 1994, 1996, Silverman & Skonie 1997). The Flanders Interaction Analysis System and its adaptations to physical education has until now been used mostly in the area of teacher education. The multiple baseline designs used in studies and statistical methods needed for analysis have given more light to "the black box" (Locke 1983, Piéron 1996, Silverman & Skonie 1997). Also the teacher educators have been more often found to be engaged in research work than before "Reflective teaching" is recognized. (Paré 1995) One of the weakness in prevalent research procedures was, as Cheffers (1990) pointed out, that replicated designs with constant research problems were neglegted. As already Flanders in 1970 noted "replication of all research projects should be promoted" (427).

2.7 A critical discussion of interaction analysis research

In spite of the encouraging results obtained with observation instruments, certain difficulties limiting their use and application, as well as the generalization of results obtained by them, are in general associated with these methods. In addition, each observation method has special problems of its own, and its further development and application depends on the extent to which these problems can be resolved. Several aspects of Flanders-type interaction analyses have been criticized on both theoretical and technical grounds.

The most obvious limitation of the Flanders system is that it measures verbal interaction, which is only a limited portion of the total classroom interaction (Heinilä 1970, Piéron 1983a). It is based on the assumption that a teacher's verbal behavior is an adequate sample of his total behavior, and that it can be observed with higher reliability than the nonverbal behavior (Amidon & Flanders 1967b). In discussing methodological problems in classroom research, Dunkin and Biddle (1974, 54) cite Flanders in identifying the crux of the problem:

> "One of the best-known series of generalizations stated about teaching is the so-called "law of two thirds" posited by Flanders... According to this "law", two thirds of the time spent in classrooms is devoted to talk, two thirds of this talking time is occupied by the teacher and two thirds of teacher talk consists of direct influence." (p. 54).

In his investigations of teaching as a stochastic process, Komulainen noted other problems associated with the use of this method. For example, the system is suited only to teaching situations where the group of pupils acts as an undifferentiated system under the direction of the teacher. In addition, this method records interaction only in the vertical direction (teacher-pupil), when the system works as an undifferentiated whole (frontal instruction). However, horizontal interaction also occurs in groups of pupils. Komulainen also pointed out that, from the standpoint of models of the instructional process, the forms of teaching are of greater importance than the problems of subject-specificity. The

social form of instructional process decisively affects the number of necessary models" (Komulainen 1971a, 19-21).

One noteworthy solution for problems of this kind in interaction analysis is provided by multidimensional parallel codings. Flanders (1967a, 1970) suggested the use of matrices of multidimensional category systems for studying interaction models of critical teaching behaviors. In analyzing other systems, he noted that each one is designed to give emphasis to a particular conceptual framework. In multidimensional systems, elements are grouped into homogenous clusters, and each cluster is given a label. The label is usually, by definition, on a higher level of abstraction than the elements making up each set. Then the relationships between clusters can be hypothesized using the shorthand labels. Finally, from these relationships generalizations can be discussed and predictions made in an effort to apply them in different situations.

Some attempts to resolve the problems inherent in interaction analysis by multidimensional coding and matrix analysis have already been discussed. Cheffers (1973) used a "blind-live" method of validating his instrument, and "outside" and "inside" criteria coded from a videotaped original sequence of events. The comparison was made by using a variance analysis technique. Since this kind of validating procedure is not strictly a laboratory experiment nor simply an experiment in natural surroundings, they are referred to as "quasi experiments" (Cooley & Lohnes 1976).

The utility of observation instruments is usually determined by indicating the value of the reliability coefficient. Scott's method has often been used for calculating reliability indices. In most cases it signifies intercoder agreement, although within-coder constancy has also been reported in one of the studies (Kemper et al. 1976). The nonparametric coefficient of concordance, Kendalls' *W*, has also been applied for assessing the reliability of various individual categories or matrices, operationalized as inter-coder agreement.

Perhaps the most critical problem is the conceptual confusion reflected in these instruments. As Medley (1987, 176) pointed out, most category systems are built around some set of assumptions about a model of teaching, which determines how the domain of behaviors is subdived in categories. This makes subdiving exsisting categories just about the only way in which a category system can be revised without destroying it. The single-dimensional systems seem to contain overlapping aspects and the categories are not mutually exclusive. This is, however, properly required if Scott's method is to be used for the calculation of the reliability index (Scott 1955). The multidimensional approach is, from the methodological point of view, more useful than single-dimensional systems. The reliability of the different dimensions must be explored and reported both separately and in combination. The overall reliability method must be supplemented by a method through which the reliability of any individual category can be determined. The level of the reliability index must also be considered.

Reliability coefficients are often based on very small samples of events. The number of observers in the reliability tests reported here has varied from two to six. Using Scott's *pi*, the values of inter-coder agreement coefficients in

Mancuso's (1972) single-dimensional system of 27 categories varied between .91 and .92. In Underwood's (1979) nine categories system, a value of .96 was reported. With this method of calculating reliability, these coefficients seem unrealistically high. In the Kemper et al. (1976) 17 category single-dimensional system, the values of within-coder agreement coefficients varied between .67 and .90. With Gasson's multidimensional system, a mean value of Scott's pi .70 for repeated inter-coder agreement tests was reported, representing the reliability of all three dimensions.

According to Flanders (1967a), a Scott's coefficient of .85 or higher is a reasonable level of performance. Dunkin and Biddle (1974) have also noted that moderately high reliability has been reported in connection with modified single-dimensional FIAC systems. Flanders (1970) has demonstrated that an increase of categories and subdivisions is likely to be related to a decrease in reliability. The same effect has been noted in the studies using multidimensional category systems. The level of .70 accepted by Gasson (1971) seems to be appropriate.

In the studies reviewed above, these instruments have been used only by the developer himself. "Inter-investigation reliability" studies are also needed before making decisions concerning the implementation of these instruments for describing objectively interaction processes, for training teachers, and for testing hypotheses concerning the relationships between context, process and product variables (see, e.g. Rosenshine & Furst 1973). A more extensive validity and reliability analysis can be demanded of the developer of an observation system intended for widespread use. In such studies it would be appropriate to use different types of reliability coefficients together, because the inadequacy of observer agreements as the sole indices of reliability has been clearly established (Komulainen 1970, McGaw et al. 1972, Medley & Mitzel 1963). It is also necessary for the user and developer of observation systems to provide an adequate sample of data in order to demonstrate that the observations obtained are indeed representative of the universe to which they are claimed to generalize (see Cronbach et al. 1972). This has been a failure in experimental process- product studies focused on teacher effectiveness in connection with physical education (Piéron 1992, 26). Piéron presented following factor correlations and group comparisons to be taken into consideration before generalizing results in this area:

Tasks

closed ⟷ open

simple ⟷ complex

Duration of teaching period

Micro ⟷ complete unit

Students involved

From Kindergarten to University level

Teachers involved

From student teachers to master teachers

Dimension of the Classes

Micro ⟷ intact

Data processing

In research connected to the validation of developed observation instrument for physical education, factors of this kind have been considered (e.g. Heinilä 1983, 1987, 1992a, 1992b, Locke 1983).

3 REVIEW OF SOME METHODOLOGICAL ISSUES RELATED TO CLASSROOM OBSERVATION

3.1 Unit of analysis

An important decision in developing a measuring instrument is the selection of the unit of analysis. Flanders (1987, 21) suggests that: "not very much has been written about how to make choices, that determine the specifity of concepts to be used, frequency of assessment, and the complexity of a model, but these choices will be greatly influence once the researcher defines a unit of behavior". The choice of the unit of analysis for the events of teaching is both a methodological and a theoretical issue. The purpose of the study, the research design, the type of data being sought, and characteristics of the observation instrument need to be considered when selecting a unit of analysis. Observation instruments differ in their units of analysis according to the teaching events chosen for study. Biddle (1967) identified the following four possibilities used in different recording instruments:

1. Arbitary unit of time	- unit based upon specific predetermined intervals of time
2. Selected naturally	- unit depend upon the onset and termination of key events
3. Phenomenal units	- indicating a 'natural' break in the sequence of classroom events
4. Analytical units	- reflecting the key concepts that are operationally defined by the investigator.

When the aim in selecting a unit of observation is to make it possible to describe the interaction inherent in different dimensions or clusters, and to preserve the sequence of events, the choice of the observation unit is a multistage process related to the rhythm of the events themselves, to the specification of the observation procedures, to the construction of the observation schedule, and to the methods used for analysis.

3.2 Selection of statistical procedures

There are a variety of studies concerned with the selection of statistical procedures. This selection process is both a methodological and theoretical issue and is related to the validity of the measuring instrument (Flanders 1970). Most investigators use a class as their statistical unit. In interaction analysis a school class is considered a social system, an indivisible holistic unit, in which the instructional process manifests itself as an interaction process in time, the structural characteristics and sequential processes of which can be described (Bales & Strodtbeck 1951, 1967, Flanders 1970, Komulainen 1971a, 1971b, 1973).

Statistical analysis produces both primary and secondary information. Category distributions and the cell frequencies of sequence matrices represent primary information. From them can be produced secondary information, such as indices, factor structures, dimensions, discriminant functions, etc. Flanders (1970) has noted that the utility of the resulting information depends a great deal on the research design, for instance, how time periods are to be combined into a single cumulative display, or how such time periods are related to the purposes of classroom teaching.

In this context some variables describe general characteristics of the teaching-learning situation and the typical progress of events, while others describe differences between teaching situations. Both types of description are needed in the development and evaluation of an observation instrument, when assessing, e.g., the construct validity or the sensitivity of the instrument.

Statistical procedures can be divided into two general types: univariate procedures and multivariate procedures. In univariate procedures a single variable is related to a single outcome, whereas in multivariate studies several variables are combined. The most common procedures are simple correlations and analysis of variance.

Observational studies most commonly use univariate procedures of analysis. The use of multivariate procedures presents serious problems in the interpretation of results and therefore they had been rarely used, as Rosenshine (1971) noted in his review of observational studies. However, these procedures can be used to evaluate the validity and reliability of the measuring instruments (see, e.g. Heinilä 1980, Komulainen 1973, Koskenniemi & Komulainen 1969, Medley 1982).

Factor analysis has been used commonly to evaluate the construct validity of the observation instrument. As Dillon and Goldstein (1984, 399) have suggested the objectives of multiple discriminant analysis are, for the most part, generalizations of those of the two-group problem. They summarized four of the objectives that are typically found in applications of multiple discriminant analysis: (1) to find linear composites of predictor variables having the property that the ratio of between-groups to within-groups variability is as large as possible, subject to the constraint that each uncovered linear composite must be uncorrelated with previously extracted composites: (2) to determine whether the group centroids are statistically different and the number of statistically significant discriminant axes (i.e., the dimensionality of the discriminant space),

(3) to successfully assign new admission officers to set up an objective criterion for admitting a student matriculation, to decide on the number of possible predictors – such as high-school average, selection variables (test scores) - the intake officer wants to use to determine which variables are most important in judging the potential success of a student (Dillon and Goldstein 1984, 398). This multivarite method has been used in connection with observational studies (Carreiro da Costa & Piéron 1990, Hanke 1980b, Heinilä 1977b, 1983, 1988, 1990, 1992a, 1992b, Yerg & Twardy 1982).

By using multidimensional observation instruments the selection of statistical procedures is also a multistage process, and the interpretation of results needs specific attention.

3.3 Problems of design

Problems of design are related to research approach, purpose and focus of the study. Studies of teaching utilize many designs, such as the observation of a single class over many class periods using many variables. In order to organize findings of research on teaching, Dunkin and Biddle (1974) devised a model that grouped variables into four large classes which they labeled *presage, context, process* and *product variables*, based on the terminology suggested earlier by Mitzel (1960) by adding context variables.

Presage variables concern teacher characteristics such as formative experiences (i.e., age, sex, etc.), teacher training experiences, and teacher properties (i.e., intelligence, motivation, etc.).

Context variables concern the environmental conditions about which the to which the teacher has to adjust, e.g., classroom, school and community contexts and pupil characteristics.

Process variables refer to the actual activities of classroom teaching, or all observable behaviors of teachers and pupils in the classroom.

Product variables concern the outcomes of teaching. The most frequently investigated product variables are subject matter learning and attitudes toward the subject, both of which involve immediate pupil 'growth' (Dunkin and Biddle, 1974).

Using these terms Dunkin and Biddle classified the designs of research on teaching into four major types: (l) field surveys, (2) presage-process experiments, (3) process-process experiments, and (4) process-product experiments. Most observation instruments designed in the early forties and fifties were aimed at determining relationships between presage and product variables, that is, teacher effectiveness. Rosenshine (1971) and Gage (1978) chose to focus on process-product relationship in their reviews of research on teaching. The validity of measuring instruments is often tested with the use of context-process and presage-process experiments. Experiments concerning process-process relationships are difficult to control (Dunkin & Biddle 1974, Flanders 1987, Komulainen 1978) because teacher behavior is complex and, in part, responsive

to pupil behavior. In so-called performance-based teacher education programs, methods of interaction analysis have been used as a tool to help identify changes of behavior and to integrate theory and practice (e.g. Hanke 1980b, Heinilä 1977b, 1992a).

Process-product experiments have proved to be fruitful in classroom observation. In experiments of this kind, events are manipulated and the effects of different classroom experiences on pupil learning or attitudes are examined. This is the kind of design normally used by Flanders in his experiments (Flanders 1970, 1987, 20). The development of paradigms in this area has led to a division of the teaching process into various component activities, which constitute independent variables, and into criteria, such as type of achievement, which are treated as dependent variables, as in the studies of Flanders (1970). This approach has been manifested in the definition of the 'technical skills' of teaching and led to the development of microteaching and highly controllable arrangements for the modification of teacher behavior (Gage 1972).

However, in a later article, "Human Interaction models", Flanders (1987, 22) noted that the results of process-product research, like those summarized by Rosenshine and Furst (1973), Dunkin and Biddle (1974), and reanalyzed by Gage (1978), presented a prodigious research effort and constituted the most objective evidence there was about teaching, but this knowledge had limited value for the improvement of classroom instruction. Based on research reviews, he pointed out that most of references to Flanders' work on interaction analysis reported the relationship between teacher "indirecteness" and student achievement, as if a simple process-product relationship was involved, however, the purpose was to investigate variation in teacher indirecteness (Flanders 1987, 464).

As part of the Finnish investigations into the instructional process (DPA Helsinki Project, Koskenniemi et al. 1974; DPA = Didactic Process Analysis), Komulainen (1978) studied the developmental changes in the interaction patterns of the DPA classes. For this purpose, he used the content x class x period design in which content and period are repeated measures. This design was limited by the fact that only variables based on unit coding could be used. As a result of this limitation, other factors, which might influence development and change, were not identified. The methodological examination was confined to the FIAC system. A mixed approach to the analysis, using both hard and soft data, was necessary in drawing conclusions and in interpreting the results and differences between classes for the DPA Helsinki Project (Komulainen & Koskenniemi 1978).

Summarizing 'Research in Sport Pedagogy' from an empirical analytical perspective Piéron and Cheffers (1988) also used the model of Mitzel (1960), adapted and modified since by many authors in the domain of classroom teaching (Dunkin & Biddle 1974, Gage 1972) as well as in physical education (e.g. Piéron 1976, 1982a, Tousignant & Brunelle 1982a, Varstala 1996). This kind of model was originally designed not only to aid understanding of the teaching process but also to enable research to be summarized (Piéron & Cheffers 1988, 3). For this purpose they modified the model by adding the concept program variables in agreement with Tousignant and Brunnelle (1982b), considering

separately purely context, student characteristics, school equipment and program variables, type of objectives, nature of content, and direction of evaluation. Program variables, according to Piéron and Cheffers (1988) "are more important and useful in the design used, because they are within the decision-making power of teacher" (6). Results of experimental studies in physical education research based on product measures or on a combination of presage and product variables have usually provided inconclusive results, as in educational research, and knowledge gathered in this domain remained inconsistent. In connection with physical education experimental studies based on product measures, e.g. learning objectives, or educational objectives were found to be less measurable in terms of finite statistics than general fitness (Piéron & Cheffers 1988, 5).

In reviewing research reports published since 1980 and 1982 on physical education in the U.S.A., Locke (1983, 286) made a fundamental distinction between research on teaching and research on teacher education and content areas of classification. Research on teaching (ROT) is research in which teacher behaviors are the independent variable and some change in student behavior is the dependent variable, whereas in research on teacher education (ROTE) research some aspects of teacher training is the independent variable and some change in teacher behavior is the dependent variable. Naturally, not all designs for ROT and ROTE in the studies reviewed were experimental, but the familiar model used adapted from Dunkin and Biddle (1974) helped also to make the distinction between the two areas of inquiry (Locke 1983, 268).

The nine categories used by Locke (1983) for the classification of ROTE-PE in physical education studies according to topic area were: (1) presage variables (trainee characteristics), (2) context variables (program characteristics), (3) content variables (what is learned), (4) process variables (learning activities), (5) product variables (trainee/teacher behaviors), (6) research (methods and management of the knowledge base), (7) change (in program elements), (8) induction (the early years), and (9) inservice (the later years) (288).

Multiple baseline designs were found to be useful in studies connected with "reflective teaching" (see Martel 1995, Paré 1995, Piéron 1994). Both quantitative and qualitative research approaches are needed in long-term, longitudinal studies by using replicated designs with a constant research problem (Cheffers 1990, Locke 1983).

Each time an instrument is developed, it should be tested for reliability and validity. Reliability and validity are not regarded as a property of the instrument but as that of measurement. The observer and classification system together form the measuring instrument. The distinction between reliability and validity is a problem in observational studies. In general, reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods. However, even though a correlation between dissimilar subtests is probably a reliability measure, it is even closer to validity (Campbell & Fiske 1959).

As stated earlier, reliability is not a property of an instrument but of measurement. It reflects the ability of the instrument to resist the effects of

chance and to provide the same measurement results in varying circumstances. The instrument itself is neither reliable nor unreliable. Reliability can be estimated only when the instrument has been used to collect data and the data have been manipulated in some way to produce scores.

In observation studies, the concept of reliability has an entirely different content and significance from what it has, for example, in psychometric testing. An *observation instrument* is a set of procedures by means of which a trained observer can record and categorize behaviors and features in a quantifiable form. It consists of a number of items, to which the observer responds in some way dependent on the behavior (or feature) he has observed (Rowley 1976). *Categorizing* in observation research typically means the placement of each time-unit into certain classes according to a pre-designed plan. Thus, when examining the reliability of a coding problem associated with the development of observation systems, the phase of categorization has to be considered. The task of the category format is to make it easier to organize the work of observers and to express the purpose of the research. On the basis of the degree of category specification and clustering, category formats can be divided into three types containing (1) a number of mutually exclusive categories, which are either unique or constitute a dimension, (2) a main system, which consists of a limited number of categories placed into separate clusters. These generally constitute a dimension based on some model of thought (Flanders 1970, Heinilä 1971, 1977b). Because the observer and classification system together form the measuring instrument, the observer becomes an additional source of error of measurement. The measurement results may be more or less reliable depending on the manner in which the instrument is used, on the subjects or features observed, on the number, skill and training of the observers, and on the observation circumstances.

Komulainen (1973, 11) has observed that "the value of results depends crucially on the accurate use of the metalanguage of the classification system in the coding process". Therefore, in examining the reliability of a coding problem associated with the development of an observation system, attention must be paid both to the quality of information utilized and, above all, to the way in which it is used in the coding process. The questions to be answered, then, are which data yield the reliability index, and, secondly, how it can be computed. Once this much is accomplished, the adequate level of reliability may be determined.

The concept of reliability is understood in various ways, and various methods have been used to determine reliability in observation studies (e.g. Dunkin & Biddle 1974, Emmer 1972, Rosenshine 1971, Rosenshine & Furst 1973). These differences in turn are due to varying research objectives and methodological solution (Medley 1971).

Within the area of classroom observation instruments, the most commonly used form of reliability measure is *observer agreement*. The agreement coefficient is usually based on whether two or more observers were similar in their tally of total events of each type, using such terms as between-observer agreement, inter-rater agreement and inter-coder agreement (Rosenshine & Furst 1973). Komulainen (1970) uses the term inter-coder agreement to emphasize the ob-

jective and mechanical nature of observation in contrast to the subjective element inherent in judgments. Bellack et al. (1966) and Flanders (1967b), among others, specify reliability only in terms of observer agreement.

A second commonly used form of reliability measure is *stability* or *coder consistency*. This term has many different meanings, but the central idea is that the coder must be capable of repeating his coding later in the same way. Roughly speaking, it refers to the constancy with which the same observer codes identical audiovisual tapes or transcripts at two different times (Rosenshine 1971).

In addition, the *consistency of the trait* to be measured received increased attention. As early as 1953, Borgatta and Bales (1953) pointed out that if common elements exist in the condition under which the behavior occurs (i.e., the task, subject or size of groups), a certain degree of consistency in the interaction pattern may be expected. They also pointed out that in observation studies the term "consistency of observed phenomena" becomes a more correct identification than "reliability of test". Therefore, indices of observer agreement should not be cited as evidence of reliability.

The problem with a series of reliability indices is that each of them measures the effect of only one or two sources of error. The range of sources of error with the multifacet concept and technique of observational procedures is large. Therefore, a major problem is to decide which sources of error in measurement are relevant. In general, the magnitude of errors is regarded as primarily dependent on the type of decisions to be made from scores, as well as on how they were collected. In constructing the theory of generalizability of scores and profiles, Cronbach et al. (1972) noted that "there is a universe of observations, any of which would have yielded a usable basis for the decisions". In connection with this theory, the question of reliability, too, revolves into a question of the accuracy of generalizations, or of generalizability. The term "universe" is applied to conditions under which the subjects (or aspects) might be observed, and the term "facet" to conditions of a certain kind. The observations and measures may be classified according to the facet, the observer, the setting in which the observation is made, etc. The facets, alone or in combinations, define the universe. The universe to which an observation is generalized depends on the practical or theoretical concern of the decision maker (Cronbach et al. 1972).

Heinilä (1974) used the term frame factors instead of the term facet in connection with the model constructed for describing the general elements of the research into the interactional process in physical education and of the research strategy. The term "frame factors" emphasizes the characteristic role that different conditions play in regulating the formation of the interaction process. The term "frame factor" will be used here as well for the same reason. The frame factors regulating the formation of the coding process used alone or in combination, define the universe of the generalizability of the results.

In the observation studies of Medley and Mitzel (1963), each observer was regarded as a source of variability in addition to the between-person variability. In this study, reliability signified the extent to which the differences between different classes were greater than differences among codings of the same class.

Medley and Mitzel used an analysis of variance for estimating the variation attributable to each facet. In this connection the variability of the object of observation was shown to be the most important source of error variance. The inadequacy of inter-observer agreement as the sole estimation of reliability was also indicated.

However, Rosenshine (1971) noted that this meaning of reliability had been regarded as "intriguing" and difficult to interpret, because it asks not only whether the coders are coding in the same way, but also whether the teachers (or classes) are different in the variables of interest. McGaw et al. (1972) refined this method by elaborating on the means for measuring differentiation in a situation where teacher behavior is expected to vary. This variance component approach is based on Cronbach's generalizability theory (Cronbach et al. 1972), which enables the researcher to discover multiple sources of error. This method has been applied, e.g., by Tavecchio et al. (1977) and Splinter (1980) to determine the reliability of the instrument PEIAC I and II constructed to measure the interaction process in physical education classes, and adapted from Flanders' FIAC (1970) by Kemper et al. (1976).

Splinter (1980) used analysis of variance in experimental meta-level research and explorations with respect to reliability of categories of the Physical Education Interaction Analysis System (PEIAS II). The experimental design used contained four male teachers of P.E., of which two were considered representative of non-directive style of teaching, the other two more directive style of teaching. For every teacher eight lessons were videotaped, four in a games-situation and four using equipment, spread over various age groups. The 32 lessons were coded by three trained observers with PEIAS II, using a computer program and a teletype keyboard adapted for the purpose. The time interval was one second. It was assumed that the irrelevant influence on the score, the unreliability of observers, and other irrelevant effects, would be small compared to relevant influence of the variable 'teaching behavior'. The various effects on the score on a given category were compared by analysis of variance (ANOVA). The results indicated that in 10 out of the 16 categories teacher effect constituted a larger part of the total variance than any of the other effects, grade, observer or error effects. The results of the second analysis made by combining categories of PEIAS II and by comparing the results obtained in the first analysis indicated that the reliability was higher. When using PEIAS II as a whole, it was argued that the instrument yielded optimal results if a category score was computed as an average over observers and grades. (Splinter 1980, 155.)

Komulainen (1970), too, in connection with a study to determine the objectivity of coding of a modified Flanders Interaction Analysis System, presented a method in which both reliability components, observer agreement and observer consistency, are taken into account. Videotaped situations were used in this study, with the two codings occurring on occasions placed at three months intervals. The definitions involved in this method were based on the assumption of the presumably high constancy of the trait to be measured. The reliability problem was not regarded as related to the permanence of various features, as in Medley and Mitzel's (1963) study, but to the dependability of the

measurement of the features (Komulainen 1970) as in McGaw et al. (1972). Komulainen (1970) determined both the within-occasion reliability (agreement) and between-occasion reliability (stability) indices, and considered the variation of the coefficients computed attributable to different "facets" (school subjects, coder pairs and coding occasions). This assessment was based on the evaluation of the quality of the measurement scale. In this connection Komulainen considered the range of the variation of Scott's coefficient to have properties similar to those of the coefficient of correlation (Cohen 1960, Komulainen 1970).

Komulainen (1970) defines *inter-coder agreement* as the similarity between the codings performed by two independent observers at the point of time T, *within-coder constancy* as a reliability indicator resulting from recategorizing from a videotape and comparing various codings done by the same person, and *between-coder constancy* as the agreement between codings of the same situation performed at different points of time. The following simplified schematic representation of a two-observer case indicates how the various agreement indices are formed:

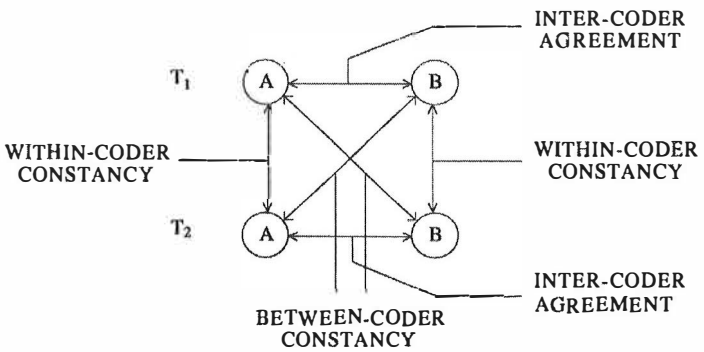


FIGURE 6 Types of various agreement indices (Source: Komulainen 1970, 6).

The method presented by Komulainen also enables the researcher to examine multiple sources of error and their characteristics, especially those caused by the coder. As Cronbach et al. (1972) and Komulainen (1973) point out the lack of reliability does not mean that the majority of classifications occur by chance. The coder's interpretation of the situation and use of the metalanguage of the classification system have been noted to be quite unique. Thus, this "source" is an additional factor causing disagreement. Komulainen has illustrated it with the following model, showing the factors contributing to reliability, the relations between these and their nature:

/ _____ / _____ / _____
reliable coder's unique chance

systematic random

(Komulainen 1973, 12)

According to Komulainen this type of error is a somewhat more important source of in the use of an observation schedule, however, since it is usually unavoidable. *Therefore, the number of coders to be used, as well as their selection and training, need to be studied in assessing the usefulness of a classification system.*

Rosenshine and Furst (1973) also addressed the same problem when comparing observation studies, in which different investigators had used the same observation instruments. They labelled this issue of reliability "inter-investigation agreement". The potential influence of observers is also closely related to the problems in determining the representativeness of coding results. If we accept that there are likely to be systematic differences between observers, then it follows that "error" variation will be greater with a team of observers than if a single observer had been used. However, by using a team the universe is broadened.

In addition, if many items are used, as in a multidimensional classification system, the "error" variation will be greater than if a single dimensional system is used, because the influence of observers will be simultaneously multiplied. Thus the increase in reliability is almost certain to be accompanied by a decrease in validity. Therefore (in this context), the classic theory of measurement errors (where reliability is regarded as a necessary but insufficient precondition of validity) is less descriptive (Cronbach et al. 1972, Komulainen 1973, Smith & Meux 1970).

The review of these issues of reliability helps us to confront the problem of multiple criterion measures. Batteries need to be produced which permit multivariate designs. In developing an observational system intended for widespread use, it is important to establish a good *within-occasion reliability* (agreement) as a necessary but not sufficient condition for stability. It is also important for its own sake when the instrument is intended to be used for feedback in connection with a performance-based teacher education program, where teacher performance is compared to a certain criterion skill used as target behavior. *Between-occasion reliability* (stability) and associated problems of representativeness are perplexing and need to be studied in this investigation, in assessing the degree of objectivity of coding. *Constancy* is also important when the observation system is intended to be used as a research tool and the object of the study is to determine if the observed variables are related to some outcome variables (Emmer 1972, Rosenshine & Furst 1973).

Unreliability may also be due to very small differences among the objects of observation on the dimension observed. It has, however, been regarded as inappropriate to delete some variables from an observation instrument even if they do not differentiate across classrooms (see, e.g., Bookhout 1967, Rosenshine & Furst 1973). It is important to take this point of view into account when developing an observational system, because it needs to be demonstrated that the observations obtained are indeed representative of the universe into which they are claimed to generalize. And as noted earlier, the universe of observations is characterized with respect to one, two or more facets (frame factors).

3.4 Estimation of reliability indices

The reliability coefficient indicates a correlation between two different uses of the same measurement. The numerical value of it can be calculated by different methods depending upon the research objectives and the nature of the material.

The reliability indices may be estimated on the unitizing level or on the distribution level. In observation studies, we are concerned with measurement events carried out by one or more persons (1, 2...n) on the same or different coding occasions (T_1, T_2, ..., T_n). For example, if two coders carry out a coding of n events independently of each other within an all inclusive and mutually exclusive group of C categories, the result is a square matrix, C x C, portrayed in Figure 7:

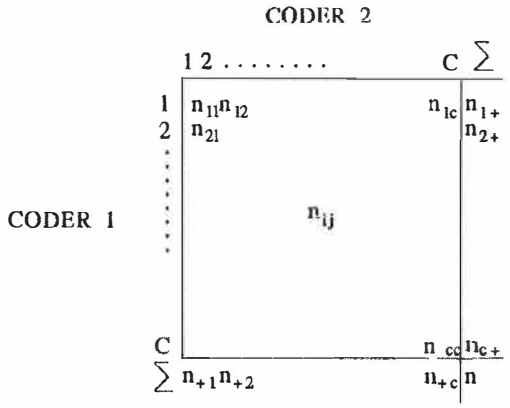


FIGURE 7 Coding occasion of two coders, with symbols used. (Komulainen 1974a, 2)

The reliability coefficients on the distribution level are based on marginal distributions ($n_1 + n_2 + ... n_c$), those on the unitizing level on diagonal frequencies ($n_{11} ... n_{cc}$) (Komulainen 1974a).

If we wish to study interactional sequences and are using matrix cell frequencies for units of analysis, reliability should be evaluated on the unitizing level. Where the nature and structure of the process are to be studied, marginal distributions may be used as the basis for reliability evaluation (Rosenshine & Furst 1973). The indices may be applied to single categories or averaged across all categories. Thus they are used to describe the overall reliability of the observation system. In the present study both systems were applied.

For estimating reliability, several indices of agreement and stability have been used, including percentage of agreement, intraclass correlation (usually the product-moment, but occasionally the rank-order coefficient) between two sets of scales. The indices based on perceived agreement give a misleading picture of reliability. For example, where few categories are involved, as in dichotomous coding, the role of chance agreement is great: disagreement in one means agreement in the other, the "errors" are compensating each other.

Therefore, in examining the objectivity of the coding of a multi-dimensional observation instrument with different numbers of categories in each cluster, there is no reason to align the reliability problem of a category system with the normal measurement of quantitative scales, where reliability is defined as the ratio of true to observed variance (Komulainen 1973).

For his Content Analysis, Scott (1955) developed an improved method for estimating reliability in the case of nominal scale coding. Scott's coefficient is a method for estimating observer reliability using any system which assigns events to mutually exclusive categories. It is applied to several categories and takes chance agreement into account by subtracting from each category the proportion of frequencies that would be expected to be in agreement by chance alone. Scott's *pi* takes into account the fact that the agreement to be expected on the basis of chance does not equal the theoretical expectation but varies according to the relative frequency of occurence of each category (P) in the sample to be analysed. The mean value of the coders' category distribution of the entire sample is computed, and from this the role of chance is computed. Scott's coefficient provides information not on individual categories, but on the mutual consistency of two coders' entire codings.

Scott's pi is virtually the only reliability index used with the Flanders Interaction Analysis Category System (FIAC). Flanders (1965) argued for this method when comparing it with the adaptation of the Chi Square proposed by Bales, and noted that Scott's method (1) is unaffected by low frequencies, (2) can be adapted to per cent figures, (3) can be estimated more rapidly in the field, and (4) is more sensitive at higher levels of reliability. Scott's coefficient *pi* used by Flanders (1965) is determined by the two formulas below:

(1) $$\pi = \frac{P_o - P_e}{1 - P_e}$$ where: P_o = observed percentage agreement
P_e = percentage agreement to be expected on the basis of chance, as obtained from (2)

(2) $$P_e = \sum_{i=1}^{k} P_i^2$$ where: P_i = the proportion of tallies falling to each category
k = the number of categories

(Scott 1955, 321-325)

In formula one, π ("pi") can be roughly interpreted as the amount by which two observers exceeded chance agreement divided by the amount by which perfect agreement exceeds chance (Flanders 1967b).

Originally Scott's coefficient was designed for computation on the unitizing level (Scott 1955). However, it is also considered applicable to reliability coefficient computation on the distribution level. Among others, Komulainen (1973) suggests, on the basis of studies on differences of individual categories between agreement coefficients on the unitizing and distribution levels, that the danger of mutually compensating errors due to the use of the frequency totals is not serious.

It can be concluded, after reviewing the possibilities for estimating reliability indices, that the criterion to be used has relevance to the

measurement scale, to the role of chance, to the level of calculation of indices, to the choice of the methods to be used for calculating the coefficient, as well as to the objectivity of coding. In addition, the problems of observer training need to be taken into account in this context.

Effective training of coders requires immediate feedback regarding how they have learned to make category discriminations. For that purpose, Flanders (1967b) developed a method, which makes it possible to estimate reliability quickly in the field by using a pocket slide rule. He modified Scott's method by converting tallies into percent figures and by developing a graphical method for estimating "P" from the size of the two largest categories (Flanders 1967b, 161-166). This method is also appropriate for the examination of the reliability of the multidimensional observation instrument.

However, coders must be given at least some training before they are able to use observation instruments. Flanders (1967b, 158) graphically describes the problem of observer training as twofold, "first, converting men into machines, and, second, keeping them in that condition while they are observing".

It was found that individuals differ in their ability to become reliable observers. In general, the persons who have become successful observers have had counselling experience, a broad background in social psychology, or experience as observers in some other system of interaction analysis. Also successful teaching experience, particularly on the elementary level, was found to be a strong predictor of a reliable observer (Flanders 1967b).

The training procedures used and the length of the training period required need to be considered. In general, the training procedures are related to the observation system used. The more complex the instrument, the more training is required before coders are able to use it reliably. For example, when using the Flanders FIAC system, the categories are first memorized. Then the training begins using a variety of tape recordings of classroom interaction, which provide unusual examples of direct or indirect influence patterns. There is an exact category distribution for each tape used. Six to ten hours of preliminary training with tapes were found necessary before coders were able to move to the second phase of training, observing in "live" classrooms. During this phase of training the presence of experienced trainers is needed.

Consistent observation by a team requires group training, discussion of common ground rules, each observer's understanding of his own unique biases, and regular meetings after training to discuss unusual categorization problems are required (Flanders 1967b, Splinter 1980).

Flanders (1970, 141) described an experiment in which the original Flanders Interaction Analysis System (FIAC) was subscripted to 22 categories. The training period for the new system consisted of 18 hours. Eighteen of nineteen reliability checks produced a Scott's coefficient between .70 and .86 with the median .79. One of the lowest coefficients (.56) occurred during a "difficult" observation and was followed by creating some ground rules, which eliminated the difficulty. When all the observations were collapsed to the original 10 categories, all reliabilities were about .05 to .10 higher.

According to Flanders (1967b, 166), a Scott coefficient of a 0.85 is a reasonable level of performance in using the 10 categories system FIAC. When

using a modification of Flanders Interaction Analysis System, with subscripts, this leads to increased categories, and in multidimensional observation systems these problems of treatment of data and the determination of the level of reliability need special attention. E.g. in using the PEIAC II observation instrument, which contains 16 categories instead of 10, the criterion was fixed at Scotts' Pi .70 by Splinter (1980, 111).

3.5 On the concept of validity

Both reliability and validity require that agreement between measures be demonstrated. A common denominator, which most types of validity concepts share in contradistinction to reliability, is that this agreement represents the convergence of independent approaches. In connection with observational studies, independence is, of course, a matter of degree. The concept of independence is usually indicated by such phrases as "outside criterion", "external variable", "criterion performance" (Campbell & Fiske 1959).

To assess validity for an instrument, one normally compares scores generated by it against some criterion measure that is known to reflect the phenomenon in which we are interested. To establish validity for an instrument when no criterion is available, Dunkin and Biddle (1974) proposed, "that we have a theory suggesting a relationship between the phenomenon and something else. If our investigation produces the predicted relationship it is then assumed that the measurement we have made was also valid" (1974, 79).

An observation instrument can be examined in terms of its face, content and construct validity. *Face validity* refers to the need to show that the instrument is somewhat "obviously" on target with its goal when compared with non-relevant instrumentation. The level of face validity depends on the quality of the category system and of the category definitions, and on whether or not the latter form a facet. The category set forms a facet if the categories provided are mutually exclusive and provide an unambiguous classification for each event that is to be coded to one or more facets (Dunkin & Biddle 1974, Foa 1965, Guttman 1954). For example, in the case of a physical education class, we might use the categories "pupils are collectively moving" and "not passive" to code examples of movement behavior. These two categories form a facet. It is also possible that the instrument may include two or more facets for which the events of teaching should be coded. Most instruments developed for research on teaching using live observation are single-faceted, such as the FIAC system. However, in studies, which can take advantage of video-recordings for more complete data, multifaceted category instruments are possible. If the observational instrument includes many facets, the possibilities of recording need to be considered.

Content validity is concerned in observational studies with the relevance of categories to the content area addressed. It measures the degree to which the instrument accurately measures what it seeks to measure in relation to content.

Content validity is commonly confirmed through outside criteria, such as a literature search, and through cognitive debate and interaction among specialists in the relevant field.

In studying teaching in physical education, Barrett (1969) operationalized the content validity of the observation system - consisting of four dimensions (movement tasks, content, guidance and student response) - by examining the following questions: (1) are all teacher-student behaviors as defined in the category system, observed in videotaped physical education lessons, and (2) can all teacher-student behaviors observed be categorized? The opinions of four experts were used to indicate the content validity of the observation system.

Construct validity can be defined as the ability of the instrument to distinguish between groups known to behave differently on the construct under study. Construct validity is not related solely to particular investigative procedures, but also to the orientation of the investigator. Once a test constructor hypothesizes that two individual groups will perform differently on his test, and designs an experiment to test this hypothesis, he is exploring its construct validity.

Criterion oriented validity involves acceptance of a set of operations as an adequate definition of whatever is to be measured. If the criterion is obtained some time after test is given it is called predictive validity.

When the researcher has no defined criterion measure of the quality with which he is concerned, and must use indirect measures, he will ordinarily test his instrument for construct validity (Safrit 1973). Here the trait of the quality underlying the test is of central importance, rather than either the test behavior or scores on the criteria (Cronbach & Meehl 1955).

Campbell and Fiske (1959) discuss convergent and discriminant validation and clarify the criteria to be found in cumulative evaluation considered jointly in the context of the multitrait-multimethod matrix. They show that to demonstrate construct validity, one needs to show that a test not only correlates highly with those variables with which it should (convergent validation), but also that it does not correlate with variables from which it should differ (discriminant validation). The multitrait-multimethod matrix is a systematic experimental design for this type of validation. To examine discriminant validity, and to estimate the relative contribution of trait and method variance, more than one trait as well as more than one method must be employed in the validational process. A careful examination of the multitrait-multimethod matrix (discriminant matrix) will indicate which concepts need sharper definition, and which concepts are poorly measured because of excessive or confusing method variance. Validity judgements based on such a matrix should be taken into account during the development of the instrument, along with the postulated relationships among them, the level of technical refinement of the methods, the relative independence of the methods and any pertinent characteristics of the samples.

Convergence means, according to Kerlinger (1973, 462), that evidence from different sources gathered in different ways all indicates the same or similar meaning of the construct, whereas discriminability means that one can empirically differentiate the construct from other constructs that may be similar, and that one can point out what is unrelated to the construct.

Splinter (1980, 154-155) investigated the construct validity of the observation instrument PEIAC II by using the data of the reliability study based on variance component approach with respect to the construct that was investigated. This was done on the basis of a number of judgements concerning a teacher's over-all teaching behavior. These assessments were considered an external criterion to determine the degree of (non) directiveness of a teacher's behavior. The behavior of the non-directive teachers was compared to that of directive teachers by means of the PEIAC II scores observed from video recorded P.E. lessons. The category rank orders were compared on the basis of the behavior occurring most frequently and occurring least frequently in both 'groups'. The results indicated that the categories were able to discriminate between directive and non-directive teaching behavior.

The increased use of technical equipment in observational studies makes the testing and evaluation of measuring instruments more efficient. Audiovisual recordings have an immediate appeal for research purposes, because they provide a wealth of details of the two media in which most classroom interaction takes place. However, measurements cannot be valid if the results are subject to error connected with the measurement situation. The effect of using an internal television system for classroom observation was studied by Komulainen (1968). It was found that the disturbing influence of the television system declined in about three weeks to a level from which it did not decrease any more (Komulainen 1968, 1970). Honigman (1970) and Cheffers (1973) used audiovisual recordings to validate their multidimensional observation instruments. Both tested the construct validity of their instruments by using the "blind-live" method, assuming that the encoded and decoded data arrays were sufficient to rival "live" or "on the spot" observation. Both found that their data descriptions were more accurate than those taken from live observations, although they did not achieve the same sensitivity as the live observers attained. A number of possible systematic biases were isolated in these studies, which may be connected with outside effects such as the technical equipment.

Flanders (1970) deals with the problem of validity in terms of models, and suggests that although no classroom interaction can ever be completely recreated or repeated, the issue of validity in coding does not rest on the impossibility of recreating what took place. Instead, it depends on whether what was encoded did in fact exist and whether the elements of the original situation are recreated in their proper perspective during the decoding process. Validity, therefore, requires accurate interpretation during both decoding and encoding.

Summary

In this chapter, the review of methodological issues related to the construction and validation of an observation instrument was presented, problems related to selection of statistical unit of analysis and statistical procedures used were discussed, problems of design in connection with research on teaching and teacher education were highlighted, and also concepts and methods used in estimation of reliability and validity were considered.

The stage is set for outlining the problems, questions, objectives and phases of the first methodological section of this study and for defining the terms needed to understand it.

4 RESEARCH PROBLEMS AND DEFINITIONS OF TERMS

The main purpose of the research program reported in this dissertation study was to develop and test a system for describing instructional processes in physical education classes. It was especially concerned with providing good descriptions of teacher-student interactions and did not attempt, for instance, to test hypotheses or to evaluate the effects of such interactions. Since there were no well-established or well-tested procedures for describing instructional procedures in physical education classes when the research program was started in the 1970s, the primary concern was to construct a feasible system and to gain information for its, subsequent implementation and application to teacher training.

Therefore, this study has a clear methodological orientation. Drawing on theories of the teaching-learning process and on available research findings, the first research task was to develop a theoretically justifiable system for describing and analysing interaction processes in the physical education classes. The second major research task was to test the ability of the developed taxonomy to yield a faithful description of the interaction, and to search for relational invariances both within instructional processes and between processes in contextual variation. The third research task was to develop a comprehensive paradigm for describing the activity forms and the formal proprieties of instructional process education in physical classes.

The new system was developed based on four main assumptions: (1) that P.E. classes differ from other classes, especially due to the greater role of the non-verbal behavior, (2) that P.E. classes vary to some extent in terms of their interaction patterns according to the type of class, (3) that the interaction patterns in P.E. classes vary according to grade level, and (4) that the interaction patterns in P.E. classes vary according to subject area in P.E.

Based on these tasks and assumptions, the present study sought to answer the following questions:

1. How can we develop an instrument that is suitable for the description of the instructional process in physical education through observation?

1.1 What is the state-of-the-art theoretical view of the instructional process?

1.2 What kinds of instruments have been used in the observation of teaching a) in general, and b) in physical education?

1.3 What does research say about the suitability of such instruments?

1.4 What should be the structure of an instrument that is designed to be used for the observation of the instructional process in physical education?

On the basis of such considerations, a system for observing and describing interaction process in P.E. classes was developed. This led to a set of further research questions:

2. How can we validate the developed system?

2.1 How reliable is the system in observing and describing interaction in physical education classes

(a) in live vs. video-recorded situations

(b) at different grade levels

(c) (c) dealing with different types of classes (subject areas)

(d) (d) in relation to the observation of other classes in P.E. using other systems, and

(e) (e) in relation to observations of other classes in other school subjects (particularly the Flanders FIAC system)?

2.2 How valid is the developed system?

(f) What are the proportions of talk vs. movement observed using the developed instrument as opposed to the proportion of talk in FIAC-type studies? Are there expected differences?

(g) Does the instrument distinguish reliably between P.E classes held by male and female teachers?

(h) Does the instrument distinguish reliably between P.E. classes held at three different grade levels?

(i) Does the instrument distinguish reliably between classes dealing with four subject matter areas?

(j) How does the empirical structure of the obtained data correspond to the theoretical construct structure?

(k) How invariable is the empirical structure across data of male and female teachers?

(l) How invariable is the empirical structure across three grade levels?

(m) How invariable is the empirical structure across four subject matter areas?

2.3 What is the power of the categories of different clusters to detect differences regarding directness – non-directness?

3 Is it possible to construct a model for describing the activity forms and instructional proprieties of P.E. process?

1.1 What are the basic elements of a possible interaction process paradigm?

1.2 What are the relationships between the basic elements and within the elements?

Definitions of terms

Before embarking on a discussion of the design and methodology of the study, we will define some of the key terms used in the study.

Teaching process

Instruction is seen as a mainly interactive process within school life, aiming at the development of the pupil's personality in accordance with educational objectives. Instruction consists of various situations, which are distinguishable from each other by the way activities are arranged. *Instruction* is a purposeful process where teaching is carried out according to internalised goals. The *form of instruction* refers to the way in which interpersonal communication is organized. It may be group work, problem solving, or programmed-teaching, and it may be either direct or indirect (Koskenniemi & Hälinen 1970).

Interaction is the basic unit of instruction. It presupposes communication between persons, and may be either indirect or direct by nature. In *interaction* two levels can be distinguished on which communication takes place, the content level and the process level. The *interaction process* is an event, which proceeds in real time. This interaction process includes the phases of orientation, activity and evaluation.

In *communication* the following components can be distinguished: *message, channels* (visual, auditive, psychomotor), *sender* and *receiver.*

The *content level* of communication refers to the subject under discussion and the material that is dealt with during teaching.
The *process level* of communication is the dual effect of individual behavior on one's self and on the other members of the group.

Observation instrument

An *observation instrument* is a set of procedures by means of which a trained observer can record and categorize behaviors and features in a quantifiable form. Two observation instruments discussed in this study are:

FIAC: The Flanders Interaction Analysis Category System *PEIAC/LH-75*: Physical Education Interaction Analysis Category System developed by Liisa Heinilä (1974). This system is based on Flander's theory (1970) and is a modification and expansion of his FIAC-System (see Heinilä 1977a). PEIAC/LH-75, II is modified version of PEIAC/LH-75 for teacher training (Heinilä 1977b).

The term *nonverbal behavior* refers to observable human behaviors, which are not expressed verbally. *Verbal behavior* refers to audible, spoken behavior. *Motor activities* are those goal-directed movement activities normally considered to be part of the subject matter of physical education such as games, gymnastics, dance, and fundamental movements. *Motor engagement time* or Academic Learning Time in Physical Education (ALT-PE) refers to percent of pupil collective activity in PE classes (see Piéron et al. 1990, 5, 19).

Direct influence refers to the teacher's verbal and nonverbal actions, which direct the pupil's actions or restrict the pupil's freedom of participation and/or initiation of activity, or criticize his behavior, or justify the teacher's authority or use of that authority. *Indirect influence* refers to those verbal statements or nonverbal actions of the teacher which encourage a student's participation and/or initiation of activity.

Categorizing means the placement of each time unit into certain classes in each cluster according to a predesigned plan.

Coding means conversion of the content of the instructional process into a form amenable to quantitative treatment.

The term *occasion* refers to the situation where trained observers are coding with a rule agreed in advance.

The term *frame factors* refers to the conditions under which the observations and codings are made (Heinilä 1974, Lundgren 1972).

Objectivity of coding signifies the degree of independence between the final results and the coder (Komulainen 1970, 1974b).

Inter-coder agreement is the similarity between the codings performed by two independent observers at a point of time (T_1, T_2, or T_3).

Within-coder constancy is the similarity between the codings done from the videotaped material at the point of time 1 (T_2) and the recoding of the same material at the point of time 2 (T_3) by the same observer.

Between-coder constancy is the agreement between codings of the same material performed by two or more different coders at different points of time (T_2-T_3).

Coding content constancy signifies the independence between the final coding results and the consistency of the coding target in inter-coder agreement, within-coder constancy and between-coder constancy.

Validity

Content validity refers to the degree to which the instrument accurately measures what it seeks to measure in relation to content.

Construct validity signifies the ability of the instrument to distinguish between groups "known" to behave differently on the construct under study.

Sensitivity is the ability of an instrument to make the discriminations required for the research problem (Cheffers 1973).

Criterion-related validity refers to the comparison of test or scale scores with one or more external variables or criterion. It is also called *predictive validity* (Kerlinger 1973, 460).

Criterion is an aspect or dimension of the quality to be evaluated, which is to be assessed and then compared with a level of this quality as basis for evaluating it (Medley 1987, 169). Evaluation of teaching may be based on one of *three distinct types of criteria*: (A) the outcomes of teaching, (B) the learning experiences of pupils (students) that teaching provides, and (C) the behavior of the teacher while teaching (Medley 1987, 169).

Convergence means that evidence from different sources gathered in different ways all indicates the same or similar meaning of the construct. (Kerlinger 1973, 462)

Internal consistency refers to accuracy with which the individual items of an instrument measure the particular construct under study – homogeneity (Guilford 1948).

Discriminalibity means that one can empirically differentiate the construct from the other constructs that may be similar, and that one can point out what is unrelated to the construct (Kerlinger 1973, 462).

Research area

ROT = Research on teaching. This is research in which teacher behaviors are the independent variable and some change in student behaviors the dependent variable.

ROTE = Research on teacher education. This is research in which some aspect of teacher training is the independent variable and some change in teacher behavior is the dependent variable (Locke 1983, 286).

5 RESEARCH DESIGN AND METHODOLOGY

5.1 Chapter overview

The procedures for constructing and testing the observation instrument are presented and discussed in this chapter. First, the general background and theoretical framework of the study will be presented followed by a description the decisions made in constructing the observation instrument. After that the focus will be on the procedures and strategies used for determining the reliability, followed by a discussion the validity of the observation instrument, and on data collection and analysis.

5.2 Construction of the observation instrument

The preliminary construction of the research model, and the observation instrument based on it, was done during the period of 1971-1973 (Heinilä 1974). The observation system developed was based on Flanders' theory (1965, 1970) and on the empirical studies of Heinilä (1970, 1971, 1974, 1977a).

The research strategy used for developing the observation instrument and analysis system is illustrated in Figure 8 (Heinilä 1977a).

In general, the decisions made in developing and analysing the system proceeded along the following lines:

1. specification of the entry situation and selection of a valid theoretical and conceptual framework,
2. the construction of mutually exclusive and exhaustive observable behavior categories derived from the conceptual framework,
3. the selection of a unit of observation and the development of adequate coding procedures for accurate system use,
4. the selection of a unit of analysis derived from the conceptual framework,
5. the determination of acceptable levels of inter-coder reliability (agreement) and intra-coder reliability (constancy levels).

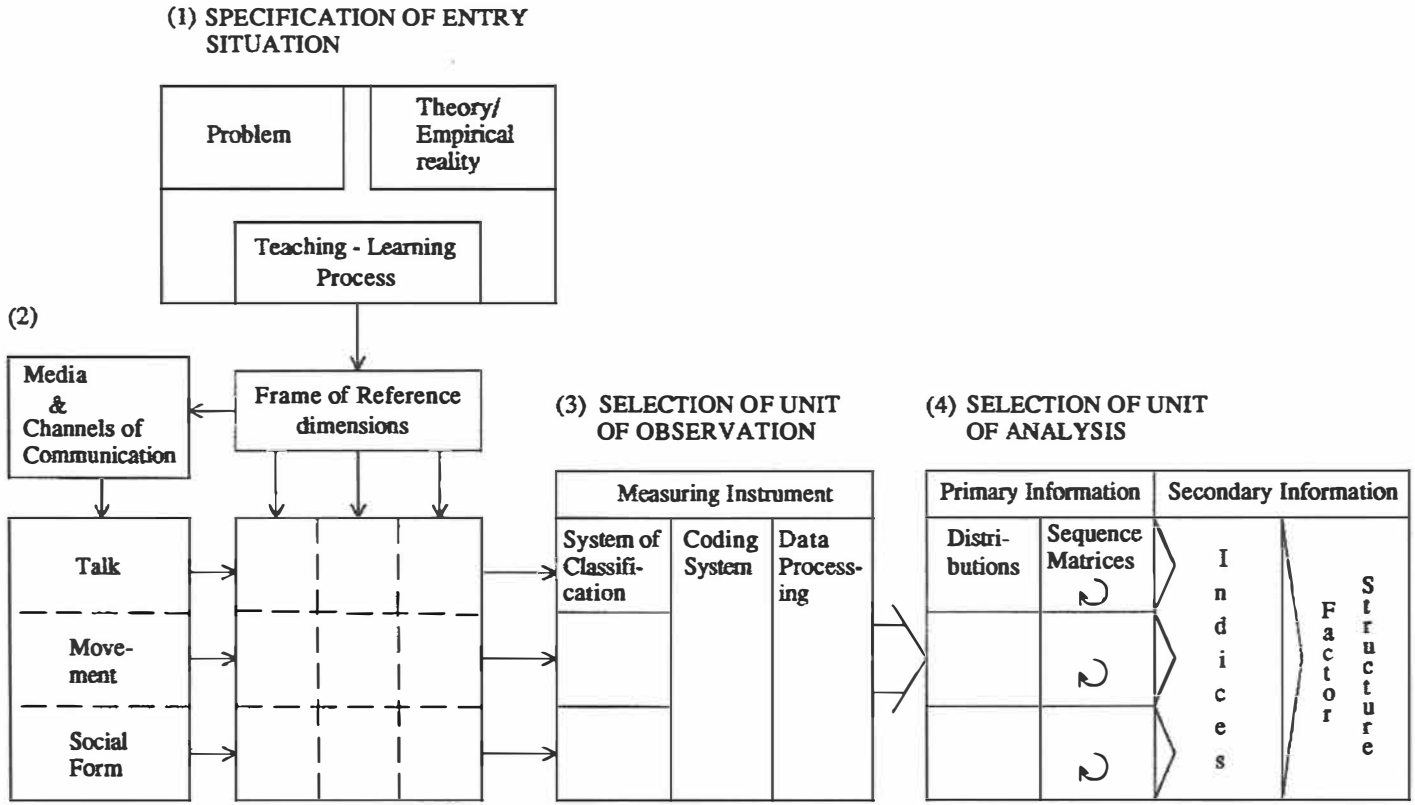


FIGURE 8 Stages and components in developing a system of analysis (PEIAC-75) (Heinilä 1977a)

A central problem was the construction of a method for the analysis of the teacher-pupil interaction in physical education in which the different factors of the interaction process and the aspects of communication could be adequately described, and so that the relevant variables of the adopted theory would be sufficiently well represented. The main task of this investigation was thus to have an adequate conception of physical education teaching, and to create an improved system for the scientific measurement, analysis, and evaluation of the physical education teaching process.

The selection of perspective was an important first step because the primary task of descriptive research is to produce an accurate record of significant real-world events. An unlimited number of objects for description and their dimensions may be identified. It is necessary to clarify which events and aspects might be significant to the development of physical education teaching, and to limit the investigation to these aspects.

Problems of content and method in the field of observation research are closely related, and should therefore be examined simultaneously. Often the measuring instrument will also include the theory, as in the classic Bales (1950) Interaction Analysis method, and the Flanders Interaction Analysis method (1970), which is perhaps the system most widely used in process research in the educational sciences. In choosing methods of this kind the researcher has not only made methodological decisions, but has also bound himself to a particular theory and group of variables. In this way the measuring instrument achieves a central significance.

Because of this close relationship between content and method, the basic functions and construct features (characteristics) of physical education teaching events are of particular importance, and must be included in the model developed for the study. Physical education teaching is an interpersonal interaction that is related to the social process of the teaching event and aims at the furthering of the pupils' personality development along the lines laid down by the educational objectives. This social interaction is located in a particular culture and way of life and has certain limitations. By taking these facts as a point of departure, the factors that become base-elements are identified as (1) the teacher and pupils, (2) the teacher-pupil interaction process, and (3) the factors regulating its constructional formation, such as, objectives, materials, and various environmental factors (Heinilä 1971, 1974, Lundgren 1972, Parsons 1968). With these base elements as a starting point, then, the following model of the interactive process of instruction was developed:

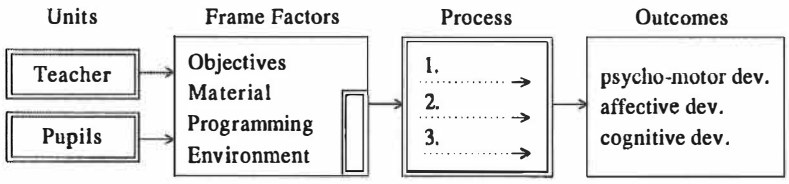


FIGURE 9 A descriptive model of the teacher-pupil interactive process in physical education (Heinilä 1974, 1977a, 221, Lundgren 1972).

It is assumed that between the elements of the model, the units, frame factors, processes and outcomes, there is a particular interrelational form which manifests itself as the selection of alternative means as the activity is directed towards the goal.

5.3 Assumptions of the study

Physical education is a situation in which the form of teaching assumes a central position. In addition, the subject matter contains a lot of affective substance and elements of creativity. A major goal of physical education is the development of pupils' independence and self-direction, i.e., a way of life characterized by physical activity and a permanent interest in physical activity (Heinilä 1971, 1974, 1976, Komiteanmietintö 1970a, 1970b).

Movement and physical exercise are typical characteristics of the interaction process in physical education. Movement communicates and movement influences. It is the goal and at the same time a means of attaining the goal. The physiological functions of exercise are realized only through movement activity. Goal-oriented teaching of physical activity is characterized by physical activity. Consequently its occurrence is an essential indicator of the teacher's mode of influence and flexibility. Therefore, the pupils' collective activity and passivity constitute an important dimension in the PEIAC/LH-75 system (see Figures 8 and 9), and at the same time represent the domain of the pupils' activity and social access.

In an active physical education situation, the social form of the participating group and the situation as a whole provide learning experiences. The social form is largely dependent on the teacher's mode of influence, which can be either a stable or transitory feature of the teaching-learning interaction process. Pupils may have different behavioral functions and roles as members of the social group. In this context, behavior refers to activities expressed by members of the group by means of verbal concepts or in symbolic terms, such as movements. Functions are forms of behavior, which are purposefully directed towards forming a group or helping it to carry out tasks (Heinilä 1971, 1974). The teacher can influence the social form of the group by the distribution of labor and responsibility within the group.

A cursory examination of the results of the pilot study of this project (Heinilä 1971) revealed the following: (a) there was a great variety of different configurations connected with the social form, division of labor and responsibility within each lesson and between lessons observed, (b) the data from 15 lessons was characterized by a diversity of content and different forms of teacher-pupil and pupil-pupil interaction, and (c) the face-to-face situation was not common. Consequently, the need for a multidimensional observation instrument was clearly indicated (Heinilä 1970, 1971).

5.4 The frame of reference

A frame of reference delimits the area of research, and defines central variables and dimensions and is determined by the research problem and a theory relevant to the exploration of the problem. It also guides the selection of the units of observation and analysis.

The balance between teacher initiation and response behavior was the focus of the observation, to be objectively measured and described in this context. This frame of reference is presented in Figures 10 and 11. It describes the theoretical and conceptual framework that was adapted for the instructional process in physical education.

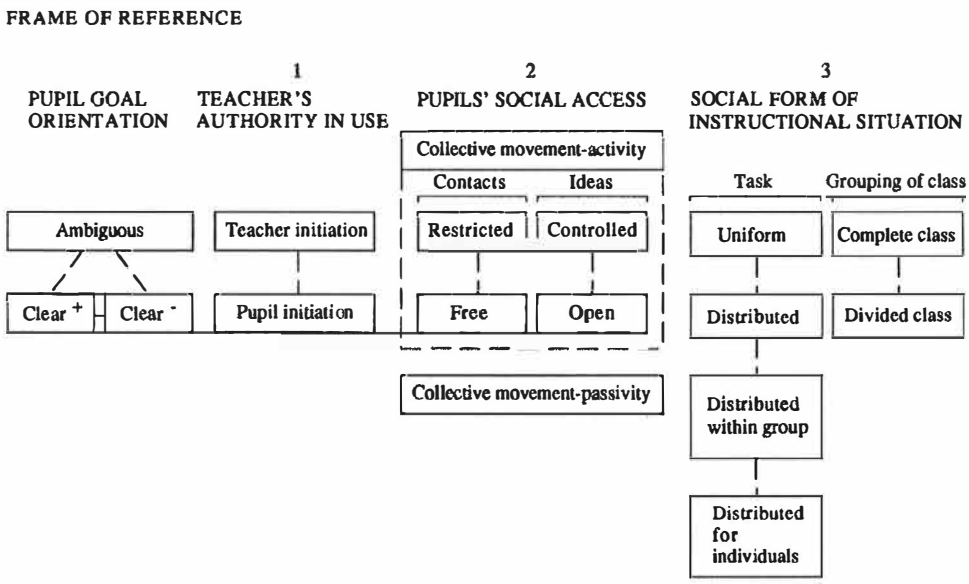


FIGURE 10 Frame of reference: Dimensions for describing the interaction process in physical education classes (Flanders 1970, 317 adapted by Heinilä 1974, 222; 1977a, 44)

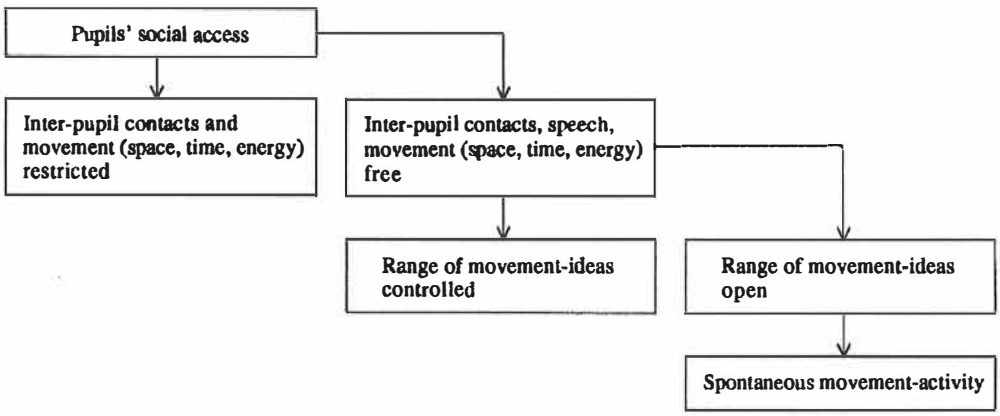


FIGURE 11 Sequence in degree of freedom of pupil's social access (Heinilä 1977a)

Given the research task of developing an observation instrument based on Flanders' theory, the first step was to adapt FIAC to better analyse and describe the interaction process in physical education classes. Flanders' theoretical model of verbal interaction was expanded by adding two aspects, which characterize interaction in P.E. classes: (1) the social access in movement activity, and (2) the social form. Accordingly, the three dimensions used to describe teacher-pupil interaction in physical education were (1) the degree of the teacher's authority, (2) the pupils' collective movement activity/passivity and social access, and (3) the social form of the instructional situation. The channels of communication and the media were taken into account in selecting the unit of observation. Thus Flanders' statement concerning his theory of changes in pupils was modified for the PEIAC/LH-75 project to read:

If...	a certain goal orientation exists (here we begin with the pupils' goal perceptions)
And...	classroom interaction is characterized by a) certain authority in use b) certain social contacts social access c) range of ideas in pupils' movement activity* (here are features of the interaction) d) and certain social form (here is division of labor and responsibility*)
Then we probably expect...	certain educational outcomes, in terms of a) pupil initiation and self-direction b) average pupil attitudes c) average subject matter achievement
	(* indicates PEIAC/LH-75 modification)

Thus, this adapted theoretical model is an attempt to explain teacher influence and changes in pupil behavior in which an intervening hypothetical mechanism is the process of goal clarification (Figure 12). Each dimension contains a certain aspect of teacher authority in use, but the channels of communication and forms are variable. Labor refers here to the behavior forms and functions that occur in the teaching situation and are similar for all members of the group or specific for individuals or groups. The execution of certain sets of functions by members of the group is referred to as roles.

Mode of teacher influence:	Dimension 1-3	Pupil goal orientation: ambiguous / clear+ / clear-		
verbal/nonverbal Direct	\ movement, social access social form /	+	±	+
verbal/nonverbal Indirect	\ movement, social access social form /	-	-	-

+ = dependence increases
- = dependence does not increase

FIGURE 12 Theoretical model for describing hypothetical mechanism in goal clarification in PEIAC/LH-75

The criterion of pupil change toward independence was believed to be an appropriate measure. The strength of this approach resides in the hope that pupil performance of required and self initiated work may be more positively identified and more precisely measured than consequent pupil change.

5.5 Selection of the unit of observation

The selection of the unit of observation is a process, which reflects both questions of principle and technique. The PEIAC/LH-75 system is based on the observation that individual classroom events are meaningful in as much as they constitute part of a sequence, and particularly as they form a sequence of interaction between teacher and pupils. Process is always in a given state. When the aim is to describe the interaction inherent in the talk, movement and the social form of the situation and to preserve the sequence of events, the choice of the observation unit is a multistage process. This is true of both the specification of the methods of observation and coding, and of the construction of the observation schedule.

In the PEIAC/LH-75 system, a unit of time occurring at given intervals was used and tallies were entered in the coding protocol at regular intervals. When category observation is based on regular time intervals, the unit of time also becomes the unit of observation. For this study, an interval of six seconds was used with triple coding. That is, each event was recorded in three different clusters. The nature, extensiveness and specificity of the unit were determined partly by the content and structure of the observation schedule and partly by the time interval.

Variables describing the sequence of events are particularly important in the study of teaching behavior since they may be related to learning outcomes. The sequence of events can be described by means of cell frequencies or indices, or by the models of behavior sequences developed from them.

The selection of the units of analysis for the description of the variables of the teaching-learning process of physical education demands careful consideration and, above all, a continuous development of research methods and their creative application.

5.6 Development of categories

The primary aim of PEIAC/LH-75 was to produce a flexible research instrument for use in describing teachers' authority in use in different physical education situations and periods. The categories of the three dimensional instrument and their respective dimensions/headings are presented in Table 2 described in the following order:

Cluster I. Teacher's talk, pupil's talk, teacher's silent activity, other
Cluster II. Pupil's collective movement activity/passivity and the social access
Cluster III. Social form (division of labor and responsibility).

Cluster I was adapted and extended from the Flanders ten-category system presented in Table 1 by making six modifications:

1. Combining FIAC-categories 1 and 2 to form the first PEIAC/LH-75 category, which thus contains acceptance, praise and encouragement by the teacher. The second category of PEIAC/LH-75 is for corrective feedback.
2. Adding to the content of the third FIAC-category (the use of the ideas), the category "movement patterns suggested by pupils".
3. Adding to the content of the fourth FIAC-category (asks questions), the category "initiates, terminates movement activity".
4. Adding to the fifth FIAC-category, the category "demonstration of movement pattern".
5. Adding to the sixth FIAC-category the category "gives direction, comments during activity (pupil expected to comply)".
6. The addition of two categories for meaningful nonverbal teacher activity: category 10. "Teacher follows pupils' activity, silent guidance" and category 11. "Teacher's silent participation in movement activity (such as dancing, playing games)".

Thus, the final categories of the PEIAC/LH-75 system are as shown in Table 2. The classifications in Cluster I were determined not only by the teacher's but also by the pupils' verbal expressions, as a result of which a certain social form was described.

In Cluster II, collective activity (categories II/1-II/4) refers to movement activity, which has a learning function. The classification was made through observation of the activity in the entire class and the degree of the pupils' freedom in movement, social contacts and range of ideas. It was used when one half of the pupils were moving. The category *spontaneous activity* (II/4) was used when pupils were allowed to move in a certain situation under the teacher's supervision and given facilities, the teacher assisting and guiding if needed. The problems were set by the pupils.

On the other hand, *movement response* (II/1-II/3) means the movement activity which was initiated by the teacher's direct or indirect actions based on his own and/or collective decisions. The term *collective movement-passivity* (II/5-II/7) indicates that pupils were not moving but were involved in other activity, which had a learning function.

Cluster III observation looks at the social form of instructional situation as a whole, which appears in the division of labor and responsibility. To classify the division of labor and responsibility, those behaviors, functions and roles, which the group members displayed during the instructional situation, were observed. Functions are behaviors directed purposefully toward building the group and toward helping it accomplish its task. These functions may be permanent of occasional, more or less conscious. The characteristic playing of certain sets of functions by group members is referred to as roles. If tasks are distributed within the group, it is the role functions that are often in question.

Decisions concerning classifications were made in all clusters on the basis of the didactic function of the activity.

PEIAC/LH-75 Categories are presented in Table 2.

5.7 Procedures in observation and coding

PEIAC/LH-75 is multidimensional and, therefore, some modification to the observation procedures used in FIAC-system was necessary. Instead of Flanders' three seconds time interval, a six second time interval was used and the triple coding produced three clusters. The dominant characteristics of the time interval were coded. Naturally, the clusters of the instrument can also be used separately, and with the first cluster, the three seconds time interval can be used, if preferred.

The procedures of observation in the PEIAC/LH-75 system were as follows: The observer placed himself where he could hear and see both the teacher and the pupils, or the video recording on the TV monitor. He observed the first five minutes from the beginning of the lesson without marking the card. The observation period was started and terminated by marking "1287" in the first and last row of the appropriate column. Then every six seconds, either on hearing the signal or by following the hands of a large clock based on top of the TV receiver, the observer decided which of the three classes of observation in the classification system the events of the previous six seconds best belonged to. The observer then wrote down the numbers selected while following the events of the next period. Thus he continued for twenty minutes making four digit markings in the appropriate row of the answer card in the six seconds columns, ten markings per minute. The chronology of the events was retained. A louder signal marked the end of a five-minute period, whereupon the observer continued marking in the first column of the row reserved for the next five minutes.

Where certain events in the observation period were unclear, an indication was made in the rows (2 vertical lines) at the beginning or end of that period and a more precise explanation was given on the right-hand edge of the card or on the back. Other features which were necessary for the later interpretation of results were indicated, for example, whether or not the class was divided, the size of the group observed, etc.

The classification time sheet (see Appendix 1) was the same as an ADP coding sheet in which the lesson material variables were coded in columns 1-8, the sequence number of the card in columns 9-10, and the observations on the teaching process within the time units in columns 11-78. Before the observation period began, the observer recorded basic information in the first ten columns of the time sheet.

TABLE 2 Physical education interaction analysis category system (PEIAC/LH-75) Heinilä 1977a

		I CLUSTER - TEACHER TALK - PUPIL TALK - SILENT TEACHER ACTIVITY category	II CLUSTER - SOCIAL ACCESS (PUPILS' COLLECTIVE MOVEMENT ACTIVITY/PASSIVITY) category	III CLUSTER - SOCIAL FORM (DIVISION OF LABOUR AND RESPONSIBILITY) category
TEACHER TALK	RESPONSE	01. Accepts, praises, encourages 02. Gives corrective feedback, directs, urges 03. Uses pupils' ideas, accepts, clarifies, develops ideas, movement, tasks suggested by pupils	PUPILS' COLLECTIVE MOVEMENT- ACTIVITY 1. Inter-pupil contacts and movement (space, time, energy) restricted; range of ideas controlled 2. Inter-pupil contacts and/or movement free; range of ideas controlled 3. Inter-pupil contacts and/or movement free; range of ideas open 4. Pupils' spontaneous activity PASSIVITY 5. Pupils follow instruction, demonstration 6. Pupils organize themselves, assists in organization 7. Pupils wait for turn	1. Complete class, uniform task 2. Divided class, uniform task 3. Divided class, differentiated task 4. Divided class, differentiated tasks distributed amongst groups & within group 5. Individual work, uniform task 6. Individual work, differentiated task
	INITIATION	04. Asks, initiates and terminates activity 05. Presents information, uses demonstration, describes, organizes pupils/material 06. Gives directions, commands during activity (pupil expected to comply) 07. Criticizes pupil behavior, rejects movement pattern		
PUPIL TALK	INIT./ RESP.	08. Answers question/clarifies demonstrates 09. Initiates speech (asking for instructions expressing own ideas, movements)		
TEACHER SILENT ACTIVITY		10. Follows pupils' activity, silent guidance 11. Silent participation in movement activity		
OTHER		12. Confused situation, uproar	8. Confuses situation, uproar	7. Other situation, confusion
The decision on classification is made on the basis of the didactic function of the activity.				

TABLE 2 (continued) I CLUSTER - TEACHER TALK - PUPIL TALK - SILENT TEACHER ACTIVITY	II CLUSTER -SOCIAL ACCESS (PUPILS' COLLECTIVE MOVEMENT ACTIVITY/PASSIVITY)	III CLUSTER -SOCIAL FORM (DIVISION OF LABOUR AND RESPONSIBILITY)
When analysing teacher's authority in use the observation is focused on teacher's and pupil's speech behaviour and the other didactic teacher activity. The decision on classification is made on the basis of the above mentioned didactic function of the teacher activity. Sequence of the actions should be retained. Catagories 1-9 The major feature of this category system lies in the analysis of initiative and response which is a characteristic of interaction between two or more individuals. To initiate, in this context, means to make the first move, to lead, to begin, to introduce an idea or concept for the first time, to express one's own will. To respond means to take action after an initiation, to counter, to amplify or react to ideas which have already been expressed, to conform or even to comply to the will expressed by others. Teacher's and pupil's initiative-response behaviour can be directed toward individuals (teacher and/or pupil), group of pupils or the entire class. The behaviour may refer either to the situation, activity or behaviour in the past, in the present or in the future. Categories 10-11 Teacher's silent, purposeful activity is classified into categories 10 and 11. In 10 his role is that of a "teacher's"; in 11 his actions are characterized by an affective identification with the pupils' actions.	Categ. 1-4 Pupil's movement responses By collective activity is meant the movement-activity which has a learning function. The decision on classification is made through observation of the activity in the entire classand the degree of pupils' freedom in movement, social contacts and range of ideas. Categ. 5-7 Other purposeful activity Collective movement-passivity means that pupils are not moving but are involved in other activity which has a learning function. Categories 1-3 Movement response means the movement-activity which is initiated by teacher's direct of indirect actions based on his own and/or collective decisions. Category 4 Activity is classified as pupils' spontaneous activity when pupils are allowed to move in a certain situation under teacher's supervision and given facilities, teacher assisting and guiding if needed. The problems are set by the pupils.	The observation is aimed at the instructional situation as a whole - at its social form which is considered to appear in division of labour and responsibility. To classify the division of labour and responsibility those behaviours, functions and roles which the group members have during the instructional situation are observed. Behaviours are actions of individual group members expressed in verbal or symbolic terms (eg movement expression). Functions are behaviours directed purposefully toward building the group and toward helping it accomplish its task. Labour: behaviours and functions, which occur in the instructional situation of P.E., may be uniform to all the pupils. Roles mean characteristic playing of certain sets of functions by group members. These functions may be permanent or occasional, more or less conscious. If the tasks are distributed within the group it is the role functions which are often in question. The desicion on classification is not only determined by the teacher's but also by the pupils' verbal expressions as a result of which a certain social form is created in the instructional situation.

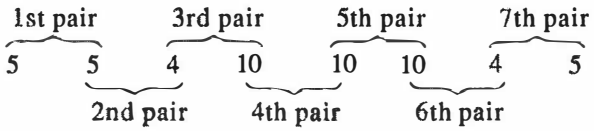
It was essential that the sequence of events be carefully preserved as it was transferred from the observers' coding sheets onto computer punch cards for the statistical processing of the material.

5.8 Matrix analysis

As stated earlier, the purpose of interaction analysis is to preserve selected aspects of interaction through observation, encoding, tabulation and then decoding. Validity in coding depends on whether what was encoded did in fact exist and whether these elements of the original situation are recreated in their proper perspective during the decoding process.

In order to preserve the elements of the original situation for accurate decoding, Flanders used a method of analysis called the *matrix analysis,* which records the sequence of events in a classroom in such a way that certain facts become readily apparent. The sequence of number codes was entered into a row/column table, or *matrix,* in which each column and each row corresponds to one of the observation categories. In the Flanders system, a 10 x 10 matrix was used (Amidon & Flanders 1967b).

The sequence of events is represented by pairs of code symbols. For example, the sequence 5,5,4,10,10,10,4,5, will read from left to right:



The first number of any pair designates the row and the second number designates the column. Note that, except for the first and last symbol, each code symbol is used twice in forming the pairs. When you use this method of pairing, there will be one less tally in the matrix than there were numbers entered in the original record (N-1), and n-1 pairs. This is a convenient way to check the tabulations in the matrix for accuracy.

In order to check for errors in recording, the first step in tabulation is to add the same number (usually the code symbol for silence or confusion) to the beginning and the end of the sequence. When a sequence of code numbers, which begins and ends with the same number, is entered into a matrix without error, the sum of each corresponding row and column will be equal. When this occurs, the matrix is said to be balanced.

In PEIAC/LH-75, the sequence of numbers of the three category clusters was entered separately by cluster, so that the first cluster forms a 12-row by 12-column matrix, the second cluster a 8 by 8 matrix, and the third cluster a 7 by 7 matrix. Separate matrices were made for each episode, with each matrix representing a single type of activity, such as class verbal/nonverbal behavior, movement activity/passivity, or social form.

5.8.1 Interpretation of PEIAC-LH-75 matrices

There are different arithmetic procedures that can be used to make comparisons between two or more matrices. They all use proportions, so that direct comparisons of numbers can be made, regardless of how long a particular observation lasted. For the PEIAC/LH-75, two general methods were used. First, all column totals were converted to a percent of the matrix total and then were calculated as ratios for which there were normative expectations. This is called a *frequency matrix*. Second, composite matrices involving thousands of tallies were converted to a common base of 1000. This is called a *millage* matrix.

Two assumptions concerning the indices were applied in this context. First, when two numbers in a matrix were added or divided, as in the calculation of a percent, the assumption was that tallying within the category system proceeded at a constant rate and each tally was presumed to be an equivalent unit. Second, as soon as an assertion, based on the matrix, was made about the classroom interaction or the social form, it was assumed that the total number of tallies and their configuration adequately represented those aspects of the original interaction which were encoded, within the limitations of the PEIAC/LH75 category system.

There were certain steps of matrix interpretation used in the PEIAC/LH-75 system, which together formed a situational setting. Adapted from the five steps of matrix interpretation used in the FIAC (Flanders 1970, 98), the first cluster consisted of five steps, the second and third clusters of four steps each.

CLUSTER I

1. Check the matrix total in order to estimate the elapsed coding time (which was usually the same for the three clusters).
2. Check the percent of teacher talk, pupil talk, silence and confusion, and teacher's silent activity, and use this information in combination with...
3....the balance of teacher response and initiation in contrast with pupil verbal and nonverbal initiation.
4. Check the initial reaction of the teacher to the termination of pupil talk, or the initiation or termination of movement activity.
5. Check the proportions of tallies to be found in "content cross" and "steady state cells" in order to estimate the rapidity of exchange, tendency toward sustained talk, toward work, and toward sustained nonverbal content emphasis.

CLUSTER II

1. Check the matrix total in order to estimate the elapsed coding time.
2. Check the percent of pupil collective movement activity and passivity, and confusion, and use this information in combination with...
3....the balance of teacher response and initiation (social access) with pupil collective movement activity.
4. Check the proportion of the tallies to be found in the "steady state cells" in order to estimate the rapidity of exchange, tendency toward sustained movement activity, and tendency toward sustained movement passivity.

CLUSTER III
1. Check the matrix total in order to estimate the elapsed coding time.
2. Check the percent of the sex differences in social forms and configurations, and use this information in combination with...
3....the balance (social form) of teacher response and initiation by division of labor and responsibility.
4. Check the proportion of the tallies to be found in the "steady state cells" in order to estimate the rapidity of exchange, and tendency of social form.

As a final step, consider emerging matrices in combination, together with certain presage and context variables (as classified according to teacher, grade level, and subject area of physical education).

5.9 The major PEAIC/LH-75 parameters and their calculation

The major PEIAC/LH-75 parameters and the formulas for their calculation are listed in Table 3. These parameters were intended to stimulate thinking about the interaction process in P.E. classes as a sequence of coded symbols and as patterns within a matrix. The indices were based on unit coding, and the statistical procedure used was category frequency matrices, with the data presented in percentages and ratios. They were computed from matrices of the three clusters of PEIAC/LH-75: indices 1-8 and 10 from the Cluster I matrix, indices 11-14 from the Cluster II matrix, and indices 15-18 from the Cluster III matrix. Index 9 was calculated by using marginal frequencies of the categories from the matrices of Clusters I and II. They can be used in interpreting and comparing PEIAC/LH-75 matrices.

It is important in comparing two or more matrices to examine the matrix totals and consider whether the sample is appropriate for the stated purposes. Matrix interpretation must then begin with certain primary features of interaction and continue with the more complex features. These primary and complex features are discussed below for each of the three clusters.

1. The proportion of *teacher talk* (TT), and...

2. ...the proportion of *pupil talk* (PT) in percent. Monopolizing talking time is one way to dominate a situation and express one's will.

3. The proportion of *teacher's sustained activity ratio* (TSAR) can be determined by calculating the percent of all tallies that lie within the 12 "steady state" cells. This ratio reflects the tendency of teacher and pupil talk, and teacher silent activity to remain in the same category for periods longer than six seconds. The higher this ratio, the less rapid is the interchange between the teacher and the pupils on the average, and the pupils may, in fact, be quite silent.

TABLE 3 PEIAC/LH-75 indices and their calculation.

No	Symbol	Name of Index	Cluster	Formulas for calculation of ratios
1	TT	Percent teacher talk	I	$\frac{01+02+03+04+05+06+07}{N_I \text{ (= row totals cluster I)}} \cdot 100$
2	PT	Percent pupil talk	I	$\frac{08+09}{N_I} \cdot 100$
3	TSAR	Teacher sustained activity ratio	I	$\frac{\text{Matrix I diagonal cells}}{N_I} \cdot 100$
4	TSGPR	Teacher silent guidance and participation ratio	I	$\frac{10+11}{01+02+03+04+05+06+07+10+11} \cdot 100$
5	TRR	Teacher response ratio	I	$\frac{01+02+03+11}{01+02+03+11+06+07} \cdot 100$
6	TQAR	Teacher question and activity initiation-termination ratio	I	$\frac{04}{04+05} \cdot 100$
7	CCR	Content emphasis ratio	I	$\frac{04+05}{N_I} \cdot 100$
8	PVIR	Pupil verbal initiation ratio	I	$\frac{09}{09+08} \cdot 100$
9	PIR	Pupil initiation ratio (verbal and nonverbal)	I, II	$\frac{09}{08+09} \cdot 100 + \frac{3+4}{1+2+3+4} \cdot 100$
10	TPR	Teacher praise ratio	I	$\frac{01}{01+07} \cdot 100$
11	PCA	Percent pupil collective activity	II	$\frac{1+2+3+4}{N_{II} \text{ (= row totals cluster II)}} \cdot 100$
12	PSUAR	Pupil sustained activity ratio	II	$\frac{\text{Matrix II diagonal cells}}{N_{II}} \cdot 100$
13	PSAR	Pupil social access ratio	II	$\frac{3+4}{1+2+3+4} \cdot 100$
14	PIOR	Pupil collective activity following instruction, organizing ratio	II	$\frac{5+6}{N_{II}} \cdot 100$
15	SGWR	Pupil social group work ratio	III	$\frac{3+4}{1+2+3+4+5+6} \cdot 100$
16	PIWR	Pupil individual work ratio	III	$\frac{5+6}{1+2+3+4+5+6} \cdot 100$
17	SFVR	Social form variability ratio	III	Number of categories used (max 6)
18	SSFR	Sustained social form ratio	III	$\frac{\text{Matrix III diagonal cells}}{N_{III}} \cdot 100$

4. The *teacher's silent guidance* and *participation ratio* (TSGPR) is defined as an index, which corresponds to the teacher's tendency to use silent guidance and participation in pupil activity as, e.g., in pupils' games or dance. The higher this ratio, the more dominant movement communication is in the interaction process.

5. The *teacher's response ratio* (TRR) is defined as an index, which corresponds to the teacher's tendency to react to the verbal and nonverbal ideas and feelings of the pupils. The formula is designed so that the index will be a percent figure, never higher than 100 and never less than zero. This ratio indicates, for example, that the teacher responded to pupil talk or movement activity more often in matrix X than in matrix Y. This index is adapted from the ID-ratio of the Flanders system (Flanders 1970, 102).

6. The *teacher question and activity initiation-termination ratio* (TQAR) is defined as an index representing the tendency of the teacher to use questions, and to initiate and terminate movement activity when guiding the more content oriented part of the situation. The TQAR is the percent of all category I/04 and I/05 statements, which are classified in category I/04.

7. The *content emphasis ratio* (CCR) is rather poorly named, since many statements in categories I/03, I/06, I/08, and I/09, as well as the teacher's silent activity categories, I/10 and I/11, are also concerned directly with content. However, the content emphasis does isolate those teacher statements, which are least likely to be involved with certain process problems, which every teacher must solve, such as presenting information or initiating and terminating movement activity.

8. The *pupil verbal initiation ratio* (PVIR) indicates what proportion of pupil talk was judged by the observer to be an act of initiation.

9. The *pupil initiation ratio* (PIR) indicates what proportion of pupil talk and movement activity was judged by the observer to be an act of initiation.

10. The *teacher praise ratio* (TPR) is defined as the tendency of the teacher to praise or integrate pupils feelings into the class discussion, or movement activity, at the moment the pupils stop talking or moving, or while they are still moving.

11. The *pupil collective activity ratio* (PCA) indicates what portion of pupil time was judged by the observer to be movement activity, which is a general feature of the interaction process in P.E. classes. When this index is average or above, it reflects the teacher's tendency to use movement activity.

12. The proportion of *pupils' sustained activity* (PSUAR) can be determined by calculating the percent of all tallies that lie within the 8 "steady state cells" of the matrix Cluster II. It corresponds to the tendency of pupil collective class time to rest in the same category for periods longer than 6 seconds. The higher the ratio, the less rapid is the interchange between the different forms of movement activity/passivity.

13. The *pupil social access ratio* (PSAR) indicates what proportion of pupil collective movement activity was judged by the observer to be a movement activity of pupil initiation. It is defined as an index, which corresponds to the teacher's tendency to use and to react to the ideas and feelings of the pupils in movement activity.

14. The *pupil collective following instruction, organizing ratio* (PIOR) indicates what proportion of pupil time was judged by the observer to be this kind of movement passivity in preparation for movement activity.

15. The *pupil social group work ratio* (SGWR) indicates what proportion of pupil time was judged by the observer to be group work based on pupil responsibility. When this ratio is average or above, it reflects the teacher's tendency to divide responsibility among groups of pupils.

16. The *pupil individual work ratio* (PIWR) is an even more sensitive index, which reflects the tendency of the teacher to delegate labor and responsibility to individual pupils, when the ratio is average or above.

17. The *social form variability ratio* (SFVR) reflects the tendency of the teacher to use different social forms and division of the labor and responsibility in the P.E. class interaction process when the ratio is average or above.

18. The *substained social form ratio* (SSFR) can be determined by calculating the percent of all tallies that lie within the 7 "steady state cells" of social form. It reflects the tendency of the teacher to divide the social form. The higher the ratio, the less often labor and responsibility divided.

5.10 Training of observers

When the PEIAC/LH-75 system is used as a research tool, it is employed by trained observers in order to collect reliable data regarding teaching behaviors as a part of a research project. Systematic and thorough training procedures are needed in order to ensure this reliability.

The observers were three men and three women holding bachelor degrees in Physical Education. Their university studies had included, in their second or third year, a 32-hour basic observer course with theory and exercises, in addition to which they received 20 hours' further training for this particular task. During the initial stages of training, the observers coded from tape scripts, audiotapes, and videotapes. The last part of the training program included discussions and illustrations of the perspective. During this period the measuring instrument was given finishing touches and preliminary experiments were made on its applicability. Ground rules for coding were developed to supplement some of the operational definitions for the clusters and categories. At the end of the training period, the inter-coder agreement was estimated by using Scott's *pi*. It was shown to have reached an adequate level (MD .89). Because reliability was controlled during the training period, controls were not applied during the study itself.

After a basic fifteen-hour observation course, the categories are memorized, and training begins with videotape recordings of interaction in the gymnasium. There should be a variety of training tapes that provide examples of different indirect or direct influence patterns, different aspects of pupils' social access in movement activity, and different social forms. Working with tapes in teams of two or more is recommended. Trainees can then start and stop

the playback to discuss each classification. Ten to fifteen hours of preliminary training with tapes is often necessary before proceeding to live situations.

Reliable observation requires consideration of the total situation being observed in order to understand the individual and collective acts and social form being classified. Trainees need to be giving ground rules in order to be consistent when choices occur. The general ground rules established by Flanders were adapted to the PEIAC/LH-75 system and applied for categorizing classroom interaction (Amidon & Flanders 1967a, 126-128).

Rule 1: When not certain in which of two or more categories a statement belongs, choose the category in Cluster I (speech) and Cluster II (movement and social access) that is numerically farthest from categories I/05, II/2 and II/5.

Rule 2: If the primary tone of the teacher's behavior has been consistently direct or consistently indirect, do not shift into the opposite classification unless a clear indication of shift is given by the teacher (in Cluster I). The same principle will be applied in Cluster II in observing forms of social access and in Cluster III in observing social forms.

Rule 3: The observer must not be overly concerned with his own biases or with the teacher's intent. Rather he must ask himself the question "What does behavior mean to the pupils as far as restriction or expansion of their freedom is concerned?"

Rule 4: If more than one category occurs during the six-second interval, choose in Cluster I the category describing interaction between the teacher and pupils. If no change occurs within six seconds, repeat that category number.

Rule 5: If a confused situation is longer than six seconds, it is recorded as 12 in Cluster I, 8 in Cluster II and 7 in Cluster III.

In general, the observation training and cluster development occurred simultaneously in this study. Observation practice revealed weaknesses in category definitions, with particular categories presenting difficult coding problems. As a result, changes in the coding system were made during the training procedure. Observers need enough training so that the mechanics of recording in three clusters does not interfere with encoding and the more common events are coded consistently. The tempo of recording must be fast enough to accomplish the purpose of the investigation. In this investigation, the training period consisted of 20 hours to guarantee the proficiency of the six observers in the use of the new three-dimensional Physical Education Interaction Analysing Category System, PEIAC/LH-75.

5.11 Research design

Observation always has a definite purpose. Before observation begins there must be a carefully prepared account of the problems the research is meant to explain. This specification will determine the selection of behavior traits, data collection, statistical analysis, and the interpretation of results. The resulting classification system can be based on 1) a theory, 2) a theoretical model, 3) an existing observational system, or 4) the results of empirical research or pilot studies.

The measurements must be directed at what we wish to measure in order to fulfil the requirement of validity. Measurement cannot be valid if the results are subject to different types of sources of error mainly associated with the measurement situation.

The measurement must also be reliable. The greater the effect of random factors on the obtained results, the less reliable the obtained data. The reliability of observational measurement is largely dependent on how objectively the person who does the classification can function. In systematic observation, the important question is how carefully the manual has specified which action should be placed in a certain category, and on the other hand, how well the person who does the classification has understood the manual. In order to verify the coders' classifications, a judge should determine, first, whether or not the classifications correspond to the manual, and, second, to what extent the classifications done by two or more persons coincide. The proposed system needs to be subjected to validation and reliability measures before it can be accepted as a feasible research instrument and as a tool to be used in teacher education.

General elements for testing the validity and reliability of the observation instrument and the research strategy used are illustrated in Figure 13.

In selecting validation procedures, one commonly wishes to know how much of the test variance is attributable to each of a number of constructs, including both the intended constructs and impurities. Factor analysis, often used to explore construct validity, leads to such a report. Since the factors are uncorrelated, the squared loadings can be interpreted directly as a percentage of the test variance (Cronbach 1971).

It is important also to address the problem of representativeness (generalizability), that is, the extent to which the sample of lessons represents the interaction-taking place in the activity classes concerned.

In the present study, of interest from the point of view of validity and sensitivity were (1) how the variables describe the structure of a given group of P.E. classes as classified by, e.g., (a) sex of teacher, (b) age level of pupils, and (c) P.E. subject area, and (2) what instructional characteristics are found when one and the same set of data is analyzed by employing a variety of techniques.

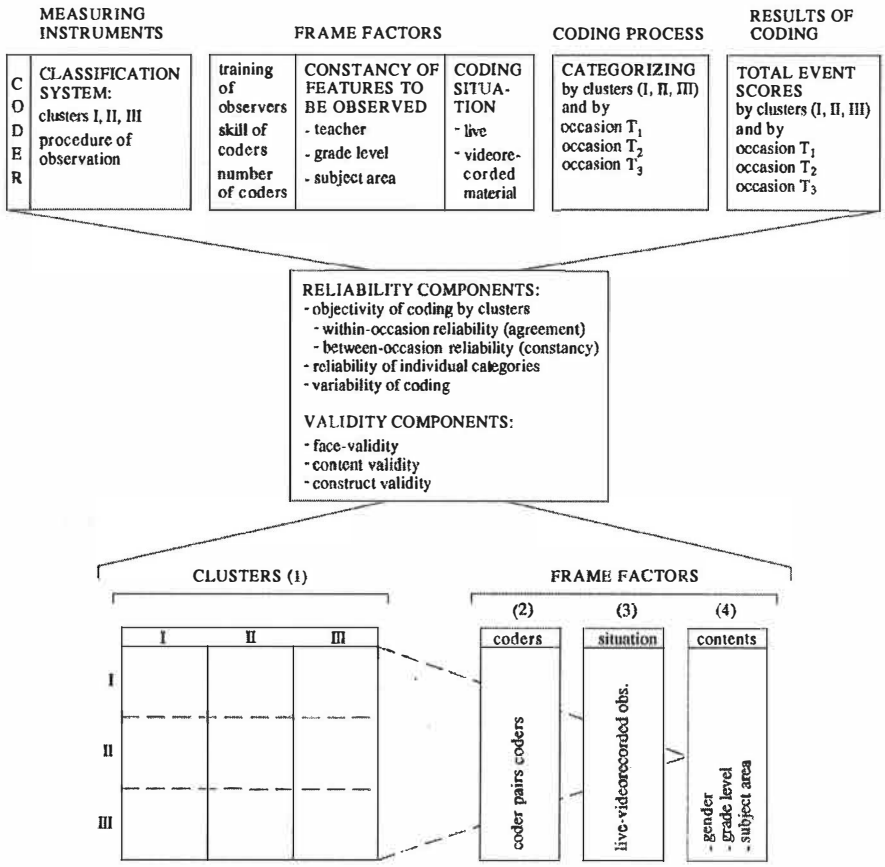


FIGURE 13 Research model: Determination of validity and reliability of observation

A major problem in developing an observation instrument is how to get adequate information for refining the classification system, and especially the rules guiding the observers so that theoretically important concepts can be measured objectively (Komulainen 1970). In evaluating the usefulness of a measuring instrument, attention must be paid both to the quality of the information available and to the way in which it is used in the coding process. The value of the results of observational studies depends crucially on the manner in which the instrument has been used in the coding process. For this reason, the present study concentrated on the objectivity of coding. In this context it signifies the degree of independence between the final results of coding and the coder himself (Komulainen 1970, 1973).

5.12 Data collection and analysis

Several different procedures were used to collect the data for determining the construct validity, sensitivity, objectivity, and reliability of coding of the observation instrument. Each of these procedures was designed to insure a controlled environment for data collection and to satisfy the requirements of a particular phase of instrument testing.

Because the study was not a laboratory experiment, nor simply an experiment in natural surroundings, the variables such as activity lessons were not chosen by means of random sampling. They were selected on the basis of theoretical considerations in an attempt to obtain a sample, which would ensure that the variables would vary in a natural way. The sampling contains the activity lessons of two teachers of different sex, with three grade levels and in four subject areas. The coding occasions included both the live situation and video recorded observation.

The activity lesson material contained different types of structures composed of the categorical elements in the three clusters of the measuring instrument, PEIAC/LH-75. These structures were intended to be either (1) common to all lessons, (2) common to a group of lessons, or (3) unique to a single lesson.

The data (Table 4) were gathered in the Faculty of Physical and Health Education at the University of Jyväskylä, in the physical education teacher training classes taught in the autumn term of 1973. The sample consisted of boys' and girls' P.E. classes at three different grade levels, covering four different subject areas for a total of 24 hours.

The observed classroom activity was recorded using the Faculty's ITV (Intern Television System (see Telama et. al., 1980 and Appendix 2). Visual recording took place with a general-purpose camera manipulated from a control room and with a manually controlled camera in the gymnasium. Audio recording took place with a general microphone and a wireless microphone of the teacher. This arrangement was intended to make the recorded material resemble the live situation as closely as possible.

The six trained coders observed the activity independently three times. They first observed the live situation (T_1), which was at the same time recorded on videotape. Then, one month later they coded from the videotaped material (T_2), and once again in another month's time from the videotapes (T_3). The time order of recorded material was randomised. Each lesson was observed for 20 minutes by the six coders, with the coding beginning five minutes after the start of each lesson. Triple coding was performed by entering four numbers on the answer sheet at six-second intervals.

TABLE 4 Research data.

Gender/ teacher pupil n		Level	n	Subject area	Coding occasion T_1	T_2	T_3
Man	12	Lower	4	Gymnastics	3	3	3
				Apparatus gymnastics	3	3	3
		Middle	4	Rhythmic movement-expression	3	3	3
		Upper	4	Ball games	3	3	3
Woman	12	Lower	4	Gymnastics	3	3	3
				Apparatus gymnastics	3	3	3
		Middle	4	Rhythmic movement-expression	3	3	3
		Upper	4	Ball games	3	3	3
Number of lessons observed					24	24	24
Number of 6-second time units					4800	4800	4800
Number of six coders total time units					28800	28800	28800

Grade levels: Lower=Grades 1-3, Middle=Grades 4-6, Upper=Grades 7-9
T_1 = live situation
T_2 = videorecorded observation 1
T_3 = videorecorded observation 2

Data analysis

The material was processed at the University of Jyväskylä Computer Center in 1974 and 1975 using the Honeywell H 1644 Time sharing system and the UNIVAC 1108/HYLPS programs D.P. and D.F. Scott's coefficients were computed with a special "Scott's" computer program designed for the purpose. The data representing the sequence of events from the six coders' coding sheets (20 coding sheets per coder for each 20-minute observation period), was recorded on computer cards. A detailed discussion of the data is presented in Chapter 6.

To determine the objectivity of the instrument, 8424 Scott's coefficients by coder pair were computed individually by cluster (I, II, III). To determine reliability, mean coefficient values and standard deviations were measured by coding occasions (T_1, T_2, T_3) for inter-coder agreement, within-coder constancy, and between-coder constancy. The variation of these component means and standard deviations was calculated by the different content situations of physical education (teacher, grade level and subject area).

The t-test was used to test the statistical significance between coder pair agreement and constancy coefficients and the same test was used to test the significance of differences between mean agreement and constancy values by cluster and by coding occasion (T_1, T_2, T_3). A one-way analysis of variance and a t-test were used to test the statistical significance of differences between mean values of inter-coder agreement, within-coder constancy, and between-coder constancy and the targets of observation (teachers, grade levels and subject areas).

The inter-coder agreement was assessed for various individual categories of the three clusters of the PEIAC/LH-75 by using the Kendall coefficient of concordance, *W* (Siegel 1956). In the statistical processing of the material, the

sub-program FORTRAN NMCC was applied. To determine the inter-coder agreement, the total percentage of frequencies, per category and per observer, and summed over the sample of 24 lessons, was ranked separately by categories of the three clusters and by occasions T_1, T_2 and T_3. A Chi Square test was used for estimating the degree of the statistical significance of the coefficients.

The intra-class correlation coefficient, based on the mean squares obtained from the six observers' percentage per category, by cluster, over a sample of 24 lessons (28 800 time units), was used to calculate the reliability of the various individual categories separately on occasions T_1, T_2 and T_3.

The starting point for a discriminant analysis for analysing variation of coding were to the six observers' score distributions of categories of the 24 lesson data (T_2), as well as the 27 categories of the three clusters of the category system.

For construct validity and sensitivity, the data of every category and cluster were analysed by analysis of variance (ANOVA), in which gender, grade level and subject area effects were analysed in terms of differences in component variance.

The scores used in calculating indices for each group were obtained from 24 lesson data (T_2) of the six observers' material (T_2) from composite matrices showing the total frequencies and average percentages of marginal frequencies. The significance of differences in means of PEIAC/LH-75 indices between frame factors (teachers, grade levels and subject areas) was computed by using the Mann-Whitney U-test.

A cumulative multivariate analysis of the factorial structure of instructional situations and grouping analysis based on the factor scores was used to analyse construct validity and sensitivity of the observation instrument. (Heinilä 1983, 1987)

In summary, the observation instrument PEIAC/LH-75 was created to enable researchers to gather valid and reliable empirical data on selected process variables of physical education classes. Such data gathering would provide a comprehensive index of teaching behavior in physical education classes upon which future teaching strategies could be based. Further, it would guide the selection and implementation of teacher training programmes if significant correlations were obtained between the scores of the student rating scale and the behaviors recorded with the observational instrument. It was assumed that with the greater number of clusters, variables and associated techniques for describing and classifying teacher-pupil behavior, the expanded instrument would be more useful and more descriptive in the physical education setting than the original (FIAC) (Heinilä 1974, 1977a).

6 RESULTS

6.1 Chapter overview

A fundamental purpose of the research reported in this dissertation was to test the reliability and validity of the observation instrument (PEIAC/LH-75) developed by the author for the description of interaction processes in physical education classes.

In this chapter the procedures for instrument testing and the results of each phase of testing are reported and discussed in three phases. Discussion will begin with a descriptive analysis of the characteristics of the observation instrument and variation of scores according to context variables in phase I. The reliability and objectivity of coding are discussed in phase II. Phase III will report on the construct validity and sensitivity of the observation instrument.

6.2 Phase I: A descriptive analysis of the observation instrument PEIAC/LH-75

In this section the characteristics of the measuring instrument and the statistical procedures used in processing the data are presented. The starting point for these analyses was the score distributions and sequence of the categories of the three clusters across class time for 24 lessons as coded by six trained observers on three separate occasions (T_1, T_2, T_3). In addition, certain frame factors such as coding situations, teachers, grade levels and P.E. subject areas are dealt with. The total coded class time for the sample was 28,800 six-second time units.

The main criterion for assessing the results was: How well does the PEIAC/LH-75 system work? The approach adopted for this study is descriptive. The data should essentially speak for itself, and is presented as directly and simply as possible. Furthermore, the discussion of the results is directed primarily at providing insights into the subtleties of the system and its

application and into the limitations of the data. The results will be presented in terms of the following four major components:

1. Describing the use of PEIAC/LH-75 in live and in video recorded observations, assessed by analysing the variation of means by categories of the three clusters as a function of the coding situation and as a control repetition coding from video recorded material (T_1-T_2, T_1-T_3, T_1-T_2).

2. Describing the instructional process by means of the categories of PEIAC/LH-75. Analysis is further divided with respect to variation as a function of teachers; grade levels and P.E. subject areas.

3. Describing the instructional process with PEIAC/LH-75 by using matrix analysis to determine general aspects of sequence and variety in the interaction process across class time by mean measures. Analysis is divided further with respect to variation as a function of teachers; grade level and P.E. subject area.

4. Describing the instructional process by means of major PEIAC/LH75 parameters and indices, presented in percentages and ratios. Analysis is divided further with respect to variation as a function of teachers; grade levels and P.E. subject areas.

6.2.1 Variation according to context variables: equipment

Describing the use of PEIAC/LH-75 in live and video recorded observations

Table 5 presents the mean measures and variability for the categories of the three clusters with respect to variation as a function of the coding situations (T_1, T_2, T_3). The data were analysed by using analysis of variance (ANOVA) in terms of differences in component variance.

The results of this analysis indicated that some categories, especially those which occurred often, were somewhat similar when coded in different situations, while the means of other categories which occurred infrequently were somewhat different for the live situation than for the video recorded observation. The variation of the means of categories number I/01 (teacher accepts, praises, encourages) and I/03 (teacher uses, develops ideas, movement, tasks, suggested by pupils), was different as a function of the coding situation and these differences in means between live and video recorded observations were statistically significant. This may be due in part to technical problems because a wireless throat microphone was not used to record the teacher's voice nor the voices of the pupils, as was done later (see Heinilä 1977b). In the live situation the aspects of teacher response behavior which are directed mostly to individuals may be easier to recognize.

It can be concluded that in general the systematic observation of physical education classes using the multidimensional category system PEIAC/LH-75 is possible with video recorded material as well as more sensitive observations in live situations.

TABLE 5 Means, standard deviations and percentages of the classtime by categories of three clusters of PEIAC/LH-75. Significance of differences in means estimated between coding occasions: T_1- T_2, T_1- T_3 and T_2- T_3 separately by clusters

Cluster			Categories	T_1 (live situation) N=24			T_2 (videorec. obs. 1.) N=24			T_3 (videorec. obs. 2) N=24			Difference df=46			Total N=72		df 2 df 69
				mean	s	%	mean	s	%	mean	s	%	T_1-T_2 t	T_1-T_3 t	T_2-T_3 t	mean	s	F
I			Teacher's talk, movement, pupils' talk, other															
	Teacher	01.	Accepts, praises	53.9	34.2	4.5	36.8	22.6	3.1	36.9	21.1	3.1	-2.03 *	-2.06 *	.02	42.5	27.5	3.25*
		02.	Gives corr. feedback	61.1	40.0	5.1	67.3	44.4	5.6	53.0	36.7	4.4	.51	-.73	-1.22	60.5	40.3	.75
		03.	Uses ideas dev. by pup.	9.0	7.2	0.8	3.8	3.9	0.3	4.3	5.2	0.4	-3.10**	-2.56 *	.41	5.7	6.0	6.32***
		04.	Asks, init., term. act.	98.2	49.3	8.2	80.8	56.4	6.7	86.2	58.2	7.2	-1.14	-.77	.33	88.4	54.5	.64
		05.	Presents inform., org.	451.1	122.8	37.6	475.6	107.1	39.6	505.3	188.0	42.1	.73	1.56	.91	477.4	116.6	1.31
		06.	Gives dir., comm.	51.9	42.8	4.3	46.1	53.4	3.8	37.8	44.9	3.1	-.41	-1.11	-.58	45.3	47.0	.53
		07.	Criticizes	15.0	18.0	1.2	9.3	12.3	0.8	9.0	12.2	0.8	-1.27	-1.36	-.11	11.1	14.5	1.31
	Pupil	08.	Answers questions	10.1	9.3	0.8	7.1	9.5	0.6	9.1	10.0	0.8	-11.14	-.39	.71	8.7	9.6	.64
		09.	Speaks spontan., init.	23.1	20.5	1.9	20.0	17.0	1.7	16.2	15.3	1.3	-.58	-1.33	-.81	19.8	17.7	.92
	Teacher	10.	Silent guidance	337.0	159.0	28.1	370.8	155.1	30.9	360.0	161.0	30.0	.75	.50	-.24	355.9	156.7	.28
		11.	Silent participation	73.3	112.1	6.1	69.8	102.8	5.8	69.8	104.1	5.8	-.1	-.11	.00	70.9	104.9	.87
	Other	12.	Confused situation	16.3	12.6	1.4	12.6	1.5	1.1	12.4	1.1	1.0	-1.45	-1.50	-.33	13.8	7.5	2.14
				1200		100.0	1200		100.0	1200		100.0				1200		
II			Pupil's collective movement activity/passivity and social access															
	Activity	1.	Contacts, ideas cont.	177.1	208.0	14.8	136.8	199.8	11.4	125.2	192.3	10.5	-.69	-.90	-.20	146.3	198.6	.44
		2.	Contacts free, ideas cont.	452.3	270.2	37.7	488.0	285.4	40.7	507.7	279.6	42.3	.44	.70	.24	482.7	275.5	.24
		3.	Contacts free, ideas open	118.6	208.5	9.9	97.0	193.9	8.1	95.6	187.4	8.0	-.37	-.40	-.03	103.7	194.3	.10
		4.	Pupils' spont. activity	7.1	18.9	0.6	5.8	18.6	0.4	4.0	9.7	0.3	-.25	-.72	-.42	5.6	16.1	.22
	Passivity	5.	Pupils follow instruction	810.7	131.3	25.9	326.3	130.6	27.2	334.9	139.2	27.9	.41	.62	.22	324.0	132.3	.20
		6.	Pupils organization	107.2	53.2	8.9	125.6	63.4	10.5	114.3	63.4	9.5	1.09	.43	-.64	115.7	58.7	.59
		7.	Pupils wait for turn	12.7	20.4	1.0	7.7	8.9	0.6	5.3	5.8	0.4	-1.10	-1.76	-1.10	8.6	13.4	1.94
	Other	8.	Confused situation	14.3	5.0	1.2	12.8	1.6	1.1	13.0	3.0	1.1	-1.36	-1.01	.36	13.4	3.5	1.19
				1200		100.0	1200		100.0	1200		100.0				1200		
III			Social form															
	Situation	1.	Complete class, uniform task	275.0	333.4	31.3	377.7	333.0	31.5	382.9	343.1	31.9	.03	.08	.05	378.5	331.8	.34
		2.	Divided class, uniform task	327.4	390.2	27.3	336.0	412.0	28.0	340.3	386.2	29.1	.07	.19	.11	337.6	331.8	.18
		3.	Divided class, different tasks	281.0	350.3	23.4	271.5	338.1	22.6	269.5	343.2	22.4	-.10	-.11	-.02	274.0	339.1	.76
		4.	Div. cl. diff. task within gr.	107.3	177.4	8.9	107.8	185.1	9.0	100.6	180.4	8.4	.01	-.13	-.14	105.3	178.5	.12
		5.	Indivudual work, unif. tasks	87.8	177.6	7.3	88.7	175.6	7.4	81.3	161.3	6.8	.02	-.13	-.15	85.9	169.2	.13
		6.	Individual work, diff. tasks	3.6	17.4	0.3	3.0	14.7	0.2	2.5	12.0	0.2	-.11	-.24	-.14	3.0	14.6	.30
		7.	Other, conf. situation	17.9	21.3	1.5	15.3	15.7	1.3	13.9	7.3	1.2	-.47	-.85	-.39	15.7	15.7	.38
				1200		100.0	1200		100.0	1200		100.0				1200		

6 observers
24 lessons
4800 6 second time units, tot. 28800 time units

* = $p < 0.05$
** = $p < 0.01$
*** = $p < 0.001$

6.2.2 Describing the instructional process by means of the categories of PEIAC/LH-75 according to contextual variation

The data (the six observers' score distribution of every category of the three clusters for the 24 lessons, 28,656 six-second time units (T2) were analyzed with respect to variation as a function of teachers and frame factors by using analysis of variance (ANOVA) in which differences between scores were estimated in terms of component variance (Table 6, 7, 8 and 9).

The score distribution clearly indicates that the teachers observed consistently emphasized their own verbal (60% of the class time) rather than nonverbal behaviors, and that most of the teacher talk was "initiation". The predominant teacher verbal behavior was "presenting information and organizing" (I/05, 39,6% of classtime). The variability of teacher verbal behavior, "silent guidance" (I/10) and "silent participation" (I/11) from class to class was high and related to pupil behavior and especially to the content of instruction, i.e., the P.E. subject area (Table 9). The variation of categories, e.g., the forms of verbal initiation behavior was related to teacher sex (Table 7). The female teacher used more "initiation and termination of activity" (I/04) and "command during activity" (I/6), which is typical of the "command technique" of women's gymnastics. The interaction on the pupils' part was mostly nonverbal (99% of the class time) and differed somewhat from class to class. Interclass differences were to a considerable degree related to certain frame variables, notably pupil variables, such as sex and age of pupils (Table 8). Pupil speech behavior was mostly initiation.

With regard to pupil nonverbal participation, operationalized as *movement activity/passivity and social access*, PEIAC/LH-75 categories clearly indicated that the interaction on the pupils' part was mostly "collective movement activity" (60% of the class time), or preparation for it by "following instruction" (II/5) or "organizing themselves" (II/6) (30% of the class time). Pupils' movement activities were response behavior, also characterized by teacher initiation as analysed by the social access categories. This was emphazised in movement activities where "inter-pupil contacts and/or movements are free but the range of ideas is controlled" (II/2) (40% of class time). The use of the pupils' own ideas in movement activity was strongly related to certain frame variables, such as the P.E. subject area (Table 9).

Variability of the social form, division of labor and responsibility, from class to class was typical (Table 6). The predominant social form (39% of the class time) was "complete class, uniform task" (III/1), which was used, e.g., in situations where pupils are following instruction. However, the use of other social forms (e.g., divided class) was also very common, with a uniform task (28% of the class time) as well as with different tasks (22% of the class time). Individual work, especially with differentiated tasks, was used rarely. The distribution of the social forms was strongly related to the content of instruction, i.e., the P.E. subject area (cf. Table 9).

TABLE 6 Physical education interaction process by variables of the PEIAC/LH-75: videorecorded material (T_2), means, standard deviations, range, percentage

	Cluster		Categories	Mean	S	Max-Min	%
I			Teacher's talk, movement, pupils' talk, other				
	Teacher	01.	Accepts, praises	36.8	22.6	19.00-0.00	3.1
		02.	Gives corr. feedback	67.3	44.4	46.00-0.00	5.6
		03.	Uses ideas dev. by pup.	3.8	3.9	9.00-0.00	0.3
		04.	Asks, init., term. act.	80.8	56.4	53.00-0.00	6.7
		05.	Presents inform., org.	475.6	107.1	126.00-27.00	39.6
		06.	Gives dir., comm.	46.1	53.4	43.00-0.00	3.8
		07.	Criticizes	9.4	12.3	12.00-0.00	0.8
	Pupil	08.	Answers questions	7.0	9.5	10.00-0.00	0.6
		09.	Speaks spontan., init.	20.0	17.8	21.00-0.00	1.7
	Teacher	10.	Silent guidance	370.8	155.1	156.00-10.00	30.9
		11.	Silent participation	69.8	102.8	66.00-0.00	5.8
	Other	12.	Confused situation	12.6	1.5	6.00-1.00	1.1
							100.0
II			Pupil's collective movement activity/passivity and social access				
	Activity	1.	Contacts, ideas cont.	136.8	199.8	142.00-0.00	11.4
		2.	Contacts free, ideas cont.	488.0	285.4	167.00-0.00	40.7
		3.	Contacts free, ideas open	97.0	193.9	129.00-0.00	8.1
		4.	Pupils' spont. activity	5.8	18.6	18.00-0.00	0.4
	Passivity	5.	Pupils follow instruction	326.3	130.6	105.00-7.00	27.2
		6.	Pupils organization	125.6	63.4	56.00-2.00	10.5
		7.	Pupils wait for turn	7.7	8.9	9.00-0.00	0.6
	Other	8.	Confused situation	12.8	1.6	7.00-1.00	1.1
							100.0
III			Social form				
	Situation	1.	Complete class, uniform task	377.7	333.0	198.00-0.00	31.5
		2.	Divided class, uniform task	336.0	412.0	198.00-0.00	28.0
		3.	Divided class, different tasks	271.5	338.1	161.00-0.00	22.6
		4.	Div. cl. diff. task within gr.	107.8	185.1	107.00-0.00	9.0
		5.	Indivudual work, unif. tasks	88.7	175.6	90.00-0.00	7.4
		6.	Individual work, diff. tasks	3.0	14.7	25.00-0.00	0.2
		7.	Other, conf. situation	15.3	15.7	19.00-2.00	1.3
							100.0

6 observers
24 lessons (20 minutes) N = 144
28 800 6 second time units

In describing the instructional process using the categories of PEIAC/LH-75, twenty-two statistically significant differences as a function of frame factors were found in the 27 categories: four between the two teachers observed, five between grade levels (related to pupil behavior), and thirteen between the different P.E. subject areas.

Of the four categories describing differences between the two teachers, two were in the area of "teacher's verbal/nonverbal behavior", and two in the area of "pupil collective movement activity/passivity". These variables appear to be related to teacher education, which is somewhat different for women than for men. They reflected the characteristics of teacher initiation behavior (i.e., command technique). The instructional process was very sensitive to different frame factors, such as pupil behavior. These differences were reflected both in

teacher response and in initiation behavior, and most clearly in categories describing pupils' initiation and response behavior.

The subject area differences were statistically significant in half of the 27 categories. In most categories describing "division of labor and responsibility" and in half of the categories describing "verbal behavior", differences were statistically significant (See Table 9). Also in three categories describing "pupils" collective activity/passivity", one finds statistically significant differences between mean scores of instructional process with different content.

These are structural characteristics of the instructional process described with the three aspects of PEIAC/LH-75. Mostly they describe general features. The results are not very reliable, however, because some of the variables were used infrequently and the number of scores was low. In the next step, an attempt was made to analyse the sequential tendencies of the instructional process.

6.2.3 Matrix analysis of sequence patterns in the instructional process variation according to context variables: gender (1), grade level (2) and subject areas (3)

The millage matrices of the three clusters computed from the same data (T_2) are presented as per cents in Table 10. The millage matrices describing the interaction process from the perspective of two teachers, three grade levels and four P.E. subject areas are presented in Tables 11, 12 and 13 respectively.

In the interpretation of the results, a flow-chart description was drawn of the matrices and the cell frequencies were used to support theoretical speculations. In this context, *instructional process* means the transition of the system from one state to another as a function of time. *Transitions* are sequence pairs with different numbers; *steady states* are sequence pairs with the same number. The concept *variety* refers to the total number of different configurations, which occur in a gymnasium. The concept *sequence* refers to how many different configurations occurred in a given time period. Decoding a matrix attempts to recreate those aspects of the original instruction which were encoded by building a description of process (see Flanders 1970, 115-120).

In a flow diagram, knowledge of the clockwise rotation of events and the differences between columns and rows are essential. The steps used in analysing the three cluster matrices are as follows:

1. Search for the highest cell frequency as the starting point, and ...
2. ...locate the event which is most likely to flow (is located) by inspecting the row which is designated by the second number in the address of the starting cell.
3. Look in the row designated by the number in the address of the cell just marked.
4. Search for the next most frequent event what will be found, as before, in the row designated by the second number in the address of the present cell.

TABLE 7 Significance of differences between means estimated for the lessons of two teachers (man-woman) (T_2); t-test

Cluster		Categories	Teacher 1. Man n = 12 $\bar{x}$	s	2. Woman n = 12 $\bar{x}$	s	T-test dif. 1-2 n = 22 t		Total n = 24 $\bar{x}$	s	F-test df=1 df = 22 f	
I		Teacher's talk, movement, pupils' talk, other										
	Teacher	01. Accepts, praises	44.9	22.4	28.7	20.6	-1.85		36.8	22.6	3.42	
		02. Gives corr. feedback	69.6	38.1	64.9	51.6	-0.25		67.3	44.4	.64	
		03. Uses ideas dev. by pup.	3.9	3.8	3.6	4.1	-0.21		3.8	3.9	.42	
		04. Asks, init., term. act.	54.6	42.4	106.9	57.9	2.52	*	80.8	56.4	6.37	*
		05. Presents inform., org.	484.0	119.9	467.1	97.2	-0.38		475.6	107.1	0.14	
		06. Gives dir., comm.	24.3	14.4	68.0	68.7	2.16	*	46.1	53.4	4.66	*
		07. Criticizes	13.3	15.6	5.4	6.4	-1.61		9.4	12.3	2.59	
	Pupil	08. Answers questions	6.1	10.7	8.0	8.6	0.48		7.0	9.5	.23	
		09. Speaks spontan., init.	26.5	17.1	13.4	14.8	-2.00		20.0	17.8	4.00	
	Teacher	10. Silent guidance	389.8	176.8	351.8	135.0	-0.59		370.8	155.1	0.35	
		11. Silent participation	70.1	95.6	69.4	113.8	-0.02		69.8	102.8	0.24	
	Other	12. Confused situation	12.3	0.5	12.8	2.0	0.97		12.6	1.5	0.94	
			1200		1200							
II		Pupil's collective movement activity/passivity and social access										
	Activity	1. Contacts, ideas cont.	59.7	95.4	213.8	247.7	2.01		136.8	199.8	4.05	
		2. Contacts free, ideas cont.	631.2	295.1	344.8	210.4	-2.80	*	488.0	285.4	7.84*	
		3. Contacts free, ideas open	95.4	232.8	98.7	156.2	0.04		97.0	193.9	0.16	
		4. Pupils' spont. activity	10.0	25.7	1.5	4.9	-1.13		5.8	16.6	1.27	
	Passivity	5. Pupils follow instruction	253.3	135.1	399.3	76.1	3.26	**	326.3	130.6	10.0	**
		6. Pupils organization	131.9	59.4	119.3	69.3	-0.48		125.6	63.4	0.23	
		7. Pupils wait for turn	5.4	5.9	10.0	10.9	1.28		7.7	3.9	1.63	
	Other	8. Confused situation	13.1	2.1	12.5	0.8	-0.90		12.8	1.6	0.80	
			1200		1200							
III		Social form										
	Situation	1. Complete class, uniform task	374.4	291.5	380.9	383.3	0.05		377.7	333.3	0.22	
		2. Divided class, uniform task	270.0	335.1	402.1	432.8	0.78		336.0	412.0	0.61	
		3. Divided class, different tasks	241.8	360.5	301.2	327.3	0.42		271.5	338.1	0.18	
		4. Div. cl. diff. task within gr.	179.3	234.1	36.2	75.3	-2.02		107.8	185.2	4.07	
		5. Indivudual work, unif. tasks	115.8	206.5	61.5	142.1	-0.75		88.7	175.6	0.56	
		6. Individual work, diff. tasks	0.1	0.3	6.0	20.8	0.99		3.0	14.7	0.97	
		7. Other, conf. situation	18.5	22.2	12.2	0.6	-0.99		15.3	15.7	0.98	
			1200		1200							

6 observers
24 lessons (20 minutes) n=114
4800 6 second time units, tot. 28800 time units

* = $p \leq 0.05$
** = $p \leq 0.01$
*** = $p \leq 0.001$

TABLE 8 Significance of differences between means estimated for the lessons of three grade levels (T_2); t-test

Cluster		Categories	Grade-levels 1. Low level n=8 $\bar{x}$	s	2. Middle level n=8 $\bar{x}$	s	3. Upper level n=8 $\bar{x}$	s	T-test dif. 1-2 df=14 t	dif. 1-3 df=14 t	dif. 2-3 df=14 t	Total n=24 $\bar{x}$	s	F-test df=2 df=21 f
I		Teachers' talk, movement, pupils' talk, other												
	Teacher	01. Accepts, praises	35.5	17.7	36.9	21.2	39.0	30.1	0.24	0.36	0.16	36.8	22.6	0.76
		02. Gives corr. feedback	59.9	30.9	82.3	53.3	59.6	47.9	1.03	-0.01	-0.89	67.3	44.4	0.66
		03. Uses ideas dev. by pup.	6.9	4.7	2.6	1.6	1.8	2.8	-2.41*	-2.64*	-0.78	3.8	3.9	5.93*
		04. Asks, init., term. Act.	91.9	57.3	80.1	66.6	70.3	49.7	-0.38	-0.81	-0.34	80.8	56.4	0.28
		05. Presents inform., org.	525.6	109.1	456.3	89.6	444.7	115.6	-1.39	-1.44	-0.22	475.6	107.1	1.38
		06. Gives dir., comm..	56.6	11.2	47.4	48.2	34.4	41.4	-0.30	-0.76	-0.58	46.2	53.4	0.33
		07. Criticizes	15.1	13.0	12.4	13.8	0.5	1.1	-0.41**	-2.43**	-2.43*	9.4	12.3	4.01*
	Pupil	08. Answers questions	15.6	12.6	3.5	3.0	2.0	1.8	-2.64**	-3.03**	-1.21	7.0	9.5	7.81***
		09. Speaks spontan., init.	35.5	18.7	16.1	9.2	8.2	8.4	-2.63*	-3.76**	-1.78	20.2	17.0	9.35***
	Teacher	10. Silent guidance	307.2	142.0	396.7	91.4	408.5	208.6	1.50	1.14	0.15	370.8	155.1	1.03
		11. Silent participation	36.9	31.6	53.5	90.6	118.9	145.6	0.49	1.56	1.08	69.8	102.8	1.48
Other		12. Confused situation	13.3	2.4	12.2	0.5	12.1	0.4	-1.14	-1.29	-0.61	12.5	1.5	1.46
			1200		1200		1200					1200		
II		Pupils' collective movement activity/passivity and social access												
	Activity	1. Contacts, ideas cont.	147.9	182.0	161.1	246.8	101.3	186.6	0.12	-0.51	-0.55	136.8	199.8	0.18
		2. Contacts free, ideas cont.	421.2	244.9	518.4	283.7	524.4	345.7	0.73	0.69	0.04	488.0	285.4	0.31
		3. Contacts free, ideas open	29.8	31.7	110.4	172.3	151.0	287.2	1.27	1.18	0.34	97.0	193.9	0.80
		4. Pupils' spont. activity	13.0	31.6	4.1	6.2	0.1	0.4	-0.78	-1.15	-1.82	5.8	18.6	1.00
	Passivity	5. Pupils follow instruction	388.1	115.2	274.5	98.4	316.4	159.9	-2.12	-1.03	0.63	326.3	130.6	1.63
		6. Pupils organization	179.1	71.8	111.1	34.6	86.6	40.3	-2.41*	-3.18**	-1.30	125.6	63.4	6.91***
		7. Pupils wait for turn	7.5	5.0	7.6	8.6	8.0	12.7	0.04	-0.10	0.07	7.7	8.9	0.62
Other		8. Confused situation	13.4	2.4	12.8	1.2	12.2	0.7	-0.67	-1.28	-1.04	12.8	1.6	1.01
			1200		1200		1200					1200		
III		Social form												
	Situation	1. Complete class, uniform tasks	315.1	255.1	485.0	431.5	332.9	305.2	0.96	0.13	-0.81	377.7	333.0	0.61
		2. Diviced class, uniform task	447.2	420.2	286.8	420.9	274.1	427.4	-0.76	-0.82	-0.06	336.0	412.0	0.42
		3. Divided class, different tasks	296.4	319.1	13.2	344.9	304.9	358.8	-0.50	0.05	0.50	271.5	338.1	0.17
		4. Div. Cl. Ciff. Task within gr.	47.8	94.5	92.5	197.5	183.0	231.9	0.58	1.53	0.84	107.8	185.1	1.12
		5. Individual work, uniform task	62.9	165.4	110.0	204.1	93.1	176.0	0.51	0.35	-0.18	88.7	175.6	0.14
		6. Individual work, diff. Tasks	9.0	25.5	0.1	0.4	0.0	0.0	-0.99	-1.00	-1.00	3.0	14.7	0.99
		7. Other, conf. situation	21.6	27.2	12.4	0.7	12.0	0.0	-0.96	-1.00	-1.43	15.3	15.7	0.96
			1200		1200		1200					1200		

6 observers
24 lessons (20 minutes) n=144
4800 6 second time units, tot. 28800 time units.

* $p \leq 0.05$
** $p \leq 0.01$
*** $p \leq 0.001$

TABLE 9 Significance of differences between means estimated for the lessons of four subject areas of P.E. (T_2); t-test

Cluster		Categories	Subject area 1. Gymnastics n=6		2. Apparatus n=6		3. Rhythmic movement n=6		4. Ball games n=6		T-test dif. 1-2 df=10	dif. 1-3 df=10	dif. 1-4 df=10	dif. 2-3 df=10	dif. 2-4 df=10	dif. 3-4 df=10	Total n=24		F-test df = 3 df=20
			$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	t	t	t	t	t	t	$\bar{x}$	s	F
I		Teacher's talk, movement, pupils' talk, other																	
	Teacher	01. Accepts, praises	47.8	21.9	53.8	16.6	21.0	14.4	24.5	20.3	0.53	-2.50 *	-1.19	-3.66 **	-2.74 *	0.34	36.8	22.6	4.72*
		02. Gives corr. feedback	55.0	16.3	127.7	42.2	39.3	23.9	47.0	21.7	3.93 **	-1.32	-0.72	-4.46 **	-4.16 **	0.58	67.3	44.4	12.92***
		03. Uses ideas dev. by pup.	3.7	3.8	3.2	3.1	6.2	5.7	2.0	1.1	-0.25	0.89	-1.02	1.12	-0.86	-1.75	3.8	3.9	1.26
		04. Asks, init., term. act.	140.9	35.4	35.5	14.5	81.7	65.5	65.0	42.3	-6.74***	-1.95	-3.37 **	1.68	1.62	-0.52	80.8	56.4	6.26***
		05. Presents inform., org.	488.8	90.6	546.1	75.6	420.3	102.6	446.9	132.1	1.19	-1.23	-0.64	-2.12 *	-1.60	0.39	475.5	107.1	1.73
		06. Gives dir., comm.	105.0	81.8	30.5	12.8	30.0	22.4	19.0	11.5	-2.20	-2.17	-2.55 *	-0.50	-1.63	-1.07	46.2	53.4	5.03***
		07. Criticizes	12.2	17.2	8.7	10.3	12.0	15.0	4.5	4.8	-0.43	-1.02	-1.05	0.45	-0.90	-1.16	9.3	12.3	0.48
	Pupil	08. Answers questions	11.8	16.1	3.8	2.5	6.5	9.9	6.0	3.7	-1.20	-0.69	-0.86	0.64	1.18	-0.12	7.1	9.5	0.73
		09. Speaks spontan., init.	20.8	21.7	28.5	18.4	15.3	12.1	15.2	15.5	0.66	-0.54	-0.52	-1.46	-1.36	-0.02	20.0	17.0	0.79
	Teacher	10. Silent guidance	235.5	79.7	328.2	64.3	384.3	105.0	535.2	179.8	2.22 *	2.77 *	3.73 **	1.12	2.65 *	1.77	370.8	155.1	7.03***
		11. Silent participation	66.0	42.9	21.8	29.5	171.2	163.8	20.0	31.0	-2.08	1.52	-2.13	2.20	-0.10	-2.22	69.8	102.8	3.95*
	Other	12. Confused situation	12.5	0.8	12.2	0.4	12.2	0.4	13.4	2.8	-0.88	-0.88	0.70	0.00	1.01	1.01	12.6	1.5	0.82
			1200		1200		1200		1200								1200		
II		Pupil's collective movement activity/passivity and social access																	
	Activity	1. Contacts, ideas cont.	311.5	283.5	6.2	11.1	43.0	85.2	186.3	158.7	-2.64 *	-2.22 *	-0.94	1.05	2.77 **	1.95	136.8	199.8	4.17*
		2. Contacts free, ideas cont.	368.9	299.8	648.0	201.8	388.8	287.3	546.3	312.8	1.89	0.12	1.00	-1.81	-0.67	0.91	488.0	285.4	1.36
		3. Contacts free, ideas open	17.8	38.5	19.9	48.6	347.9	261.2	2.7	5.1	0.08	3.06 *	-0.96	3.02 *	-0.86	-3.24 **	7.0	193.9	9.32***
		4. Pupils' spont. activity	1.3	2.4	15.3	37.1	5.8	6.8	0.5	0.8	0.92	1.53	-0.80	-0.62	-0.98	-1.91	5.8	18.6	0.78
	Passivity	5. Pupils follow instruction	378.0	100.4	347.3	159.0	283.2	115.8	296.8	150.7	-0.40	-1.52	-1.10	-0.80	-0.56	0.18	326.3	130.6	0.65
		6. Pupils organization	103.2	49.4	148.0	89.5	102.2	54.5	149.2	50.6	1.07	-0.03	1.59	-1.07	0.03	1.55	125.6	63.4	1.06
		7. Pupils wait for turn	6.0	5.1	2.5	2.8	16.8	12.4	5.5	5.9	-1.47	1.98	-0.16	2.76	1.13	-2.02	7.7	8.9	4.24*
	Other	8. Confused situation	13.3	2.8	12.8	1.0	12.3	0.8	12.7	1.2	-0.41	-0.84	-0.53	-0.96	-0.26	0.56	12.8	1.6	0.38
			1200		1200		1200		1200								1200		
III		Social form																	
	Situation	1. Complete class, uniform task	699.8	444.0	327.5	124.7	407.7	207.3	75.7	127.0	-1.98	-1.46	-3.31 **	0.81	-3.45 **	-3.34 **	377.7	333.0	5.83***
		2. Divided class, uniform task	315.0	345.0	42.4	92.0	179.5	206.6	807.3	459.0	-1.87	-0.83	2.10	1.43	4.00 **	3.06 *	336.0	412.0	7.00***
		3. Divided class, different tasks	52.2	127.8	760.9	154.1	174.5	244.7	98.5	183.0	8.67 ***	1.09	0.51	-4.97 ***	-6.78 ***	-0.61	271.5	338.1	19.59***
		4. Div. cl. diff. task within gr.	120.8	134.2	41.8	75.3	64.2	157.2	204.2	300.6	-1.26	-0.67	0.62	0.31	1.28	1.01	107.8	185.1	0.90
		5. Indivudual work, unif. tasks	0.0	0.0	2.7	6.5	349.7	180.9	2.3	5.7	1.00	4.73 ***	1.00	4.75***	-0.09	-4.70 ***	88.7	175.6	22.15***
		6. Individual work, diff. tasks	0.0	0.0	0.0	0.0	12.1	29.3	0.0	0.0	0.00	1.02	0.00	-1.02	0.00	-1.02	3.0	14.7	1.03
		7. Other, conf. situation	12.2	0.4	24.8	31.4	12.3	0.9	12.0	0.0	0.99	0.45	-1.00	-0.97	-1.00	-1.00	15.3	15.7	0.97
			1200		1200		1200		1200								1200		

6 observers
24 lessons (20 minutes) n =114
4800 6 second time units, tot. 28800 time units

* = $p \leq 0.05$
** = $p \leq 0.01$
*** = $p \leq 0.001$

The flow diagram can be used to help clarify the sequence and to make the matrix display more understandable. Each cell of the interaction matrices and millage matrices indicates how many times in general the system has shifted from the state represented in the row to a state represented in the column in question.

These transition frequencies were denoted by decoding the matrices in terms of patterns. Of particular interest was the number of different configuration pairs, which occurred in general in the 24 P.E. lessons, and the total number, or variety, of different configurations in the matrices of the three clusters.

There was a great variability between the clusters of transition cells and steady state cells. On the average, 50% of all sequence pairs in the diagonal in the first cluster were in the steady state cells, more than 80% in the second cluster, and 90% in the third cluster. Thus, the tempo of transition was quite different for these different aspects. The critical decisions made by the teacher are thus strongly related to the time factor. In the first cluster, the "teaching" (5-5) and "silent guidance" (10-10) categories contain the highest percent of scores, more than half of which are in the steady state cells. The transitions in the other categories are not so strongly centralized to these cells (see Table 10).

In the second cluster, the most dominant steady state cells are "activity 2" (2-2) and "following instruction" (5-5), with more than 90% of the transitions in these categories found in these cells. Also in the third cluster, more than 95% of transitions are in the steady state cells.

In these situational settings, the critical teaching behavior is analyzed by observing critical transitions, i.e., sequence pairs with different numbers. It is probable that, in Cluster I, the most important decisions of the teacher occur in certain rows (nine and ten) and columns (one through seven). The tallies in these cells represent the first verbal reaction of the teacher at the moment when a student stops talking or moving. In Cluster II, the tallies in the cells formed by the intersection of rows three and four and columns five to eight represent the first collective passive behavior after pupils' collective activity in which pupils were initiative. In the third cluster, all tallies in the cells formed by rows three, four and five and column one represent the reaction of the teacher to direct the complete class and to make decisions connected to the next transition concerning division of labor and responsibility.

In the first cluster, one distinguishes four different patterns representing the teacher's verbal/nonverbal critical behavior. The most dominant pattern is the "silent guidance, a long teaching" pattern (10-10, 10-5, 5-5, 5-10). The second pattern is "silent guidance" and "stopping activity, teaching - starting activity, a short drill" (10-10, 10-4, 4-5, 5-5). In the third pattern, "command, teaching during activity" (6-6, 6-5, 5-5, 5-6) is found. The fourth critical sequence pattern is "silent guidance, corrective feedback, silent guidance" (10-10, 10-2, 2-2, 2-10). In general, teacher verbal initiation was a dominating characteristic, but one could also recognize the use of patterns describing teacher response behavior.

TABLE 10 Millage matrices for episodes by category with transition cells, steady state cells and percentage: videorecorded material (T_2)

Categories

Teacher

Teachers'talk, movement, pupils talk, other

01. Accepts, praises
02. Gives corr. feedback
03. Uses ideas dev. by pup.
04. Asks, init., term. act.
05. Presents inform., org.
06. Gives dir., comm.

Pupil

07. Criticizes
08. Answers questions

Teacher

09. Speaks spontan., init.
10. Silent guidance

Other

11. Silent participation
12. Confused situation

CAT	1	2	3	4	5	6	7	8	9	10	11	12	
1 *	1	3	0	3	11	0	0	0	0	6	0	0	
2 *	3	9	0	3	13	1	0	0	0	20	2	0	
3 *	0	0	0	0	1	0	0	0	0	1	0	0	
4 *	1	3	0	3	25	5	0	5	1	18	2	0	
5 *	6	9	0	30	272	10	2	0	8	46	5	2	
6 *	1	1	0	3	11	13	0	0	0	4	0	0	
77 *	0	0	0	0	3	0	0	0	0	1	0	0	
8 *	0	0	0	1	2	0	0	0	0	0	0	0	
9 *	1	2	1	0	7	0	0	0	0	2	0	0	
10 *	11	24	0	18	42	4	1	0	3	202	1	1	
11 *	1	2	0	2	5	0	0	0	0	1	43	0	
12 *	0	0	0	0	1	0	0	0	0	2	0	0	
tot	30	56	3	67	398	38	7	5	16	310	58	5	N=28656
%	3.1	5.6	0.3	6.7	39.6	3.8	0.8	0.6	1.7	30.9	5.8	1.1	

Activity

Pupil's collective movement activity/ passivity and social access

1. Contacts, ideas cont.
2. Contacts free, ideas cont.
3. Contacts free, ideas open
4. Pupils' spont. activity

Passivity

5. Pupils follow instruction
6. Pupils organization
7. Pupils wait for turn

Other

8. Confused situation

CAT	1	2	3	4	5	6	7	8	
1 *	97	1	0	0	13	1	0	0	
2 *	1	373	0	0	21	8	0	2	
3 *	0	0	76	0	2	1	0	0	
4 *	0	0	0	3	0	0	0	0	
5 *	12	23	2	0	213	16	1	1	
6 *	1	6	0	0	18	76	0	0	
7 *	1	1	0	0	1	0	2	0	
8 *	0	2	0	0	1	0	0	0	
tot	114	408	81	4	273	105	6	5	N=28656
%	11.4	40.7	8.1	0.4	27.2	10.5	0.6	1.1	

Situation

Social form

1. Complete class, uniform task
2. Divided class, uniform task
3. Divided class, different tasks
4. Div. cl. diff. task within gr.
5. Individual work, unif. tasks
6. Individual work, diff. tasks
7. Other, conf. situation

CAT	1	2	3	4	5	6	7	
1 *	309	2	1	0	1	0	1	
2 *	1	276	1	0	0	0	1	
3 *	0	0	224	0	0	0	1	
4 *	1	0	0	67	0	0	0	
5 *	1	0	0	0	72	0	0	
6 *	0	0	0	0	0	2	0	
7 *	2	0	0	1	0	0	2	
tot	316	281	227	90	74	2	7	N=28656
%	31.5	28.0	22.6	9.0	7.4	0.2	1.3	

6 observers
24 lessons (20 minutes) N = 144
28 800 6 second time units

In the second cluster matrix, the variety of different configurations describing pupils' participation was not as great. In the clockwise flow, we can distinguish the most dominant pattern, a long "pupils' movement activity" period with "inter-pupil contacts and/or movement free, range of ideas in movement activity controlled, instruction following", pattern (2-2, 2-5, 5-5, 2-5). In the second orbit, a "pupils' movement activity with total control, instruction following, organizing" pattern (1-1, 1-5, 5-6, 6-6) is found. The third critical sequence pattern is "pupils' collective activity with inter-pupil contacts free and range of ideas open, instruction following" (3-3, 3-5, 5-5, 5-3), and "pupils' spontaneous activity, pupils organize themselves, pupils follow instruction" (4-4, 4-6, 6-6, 6-5).

In the matrix of the third cluster, describing the flow as different social forms used in classes, observations are centralized in the steady state cells (90 %), and the variety of different configurations is low compared with the other clusters. The most dominant sequence pattern is the use of "complete class with uniform task, divided class with uniform task" (1-1, 1-2, 2-2, 2-1). The first critical sequence pattern is "differentiated tasks, complete class, uniform task" (3-3, 3-1, 1-1, 1-3), the second, 4-4, 4-1, 1-1, 1-4, the third, 5-5, 5-1, 1-1, 1-5, and the fourth, 5-5, 5-2, 2-2, 2-5. Thus the sequence patterns describe mostly teaching for all, then division of labor and responsibility in different forms. In describing the flow of critical sequence patterns, such as in the cells formed by rows 3-4 and columns 1 and 2, "divided class, differentiated tasks" are distributed amongst groups and within groups. In row 5, "individual work, uniform tasks", cell 5-5, the sequence, the number of different configuration pairs and the variability seem to be higher than with other, more direct social forms. The situation is thus more variable and non-directive. However, in general, the critical teaching behavior described by the cell frequencies was characterised by directness in this sample.

(1) Variation of sequence across class time as function of gender

The millage matrices by clusters computed from the scores of 12 lessons for each of male and female teachers rated by six observers (T2), each containing 14,328 six-second time units are presented in Table 11.

The dominant critical sequence pattern in the first cluster matrix for the male teacher is "silent guidance, present information, silent guidance" (10-10, 10-5, 5-5, 5-10), whereas for the female teacher it is a "silent guidance, terminates activity, present information, initiation of activity" pattern (10-4, 4-5, 5-5, 5-4). The second different critical pattern for the woman teacher is "teacher gives direction, commends during activity, gives information, follows pupils' activity, silent" (6-6, 6-5, 5-4, 4-10), and for the man "pupils' verbal initiation, teaching" (9-9, 9-5, 5-5, 5-9) and "silent participation, teaching" (11-11, 11-5, 5-5, 5-11) patterns. The variety of transitions and configurations was greater for the woman teacher as described in the sequence matrix of the first cluster.

In the second cluster, the dominating critical sequence pattern for the man teacher was "pupils' collective movement activity where inter-pupil contacts and/or movement free range of ideas controlled, pupils follow instruction" (2-2,

2-5, 5-5, 5-2). For the woman teacher the pattern was "pupils moving collectively, inter-pupils follow instruction, range of ideas controlled" (1-1, 1-5, 5-5, 5-1, 1-1, 1-5). The variety of configurations for the woman teacher was greater than for the man teacher.

In the third cluster, the most dominant sequence pattern for both teachers was "complete class, uniform task, divided class, uniform task" (l-l, 1-2, 2-2, 2-1). For the man teacher, the critical sequence pattern 4-4, 4-1, 1-1, 1-2 was common as was the pattern 2-2, 2-5, 5-5, 5-2. In general, the variety of social forms configurations and non-directiveness reflected through division of labor and responsibility were higher for the man teacher.

In general, the behavior of the two teachers of the sample was quite homogeneous. It was evident that they were rather flexible. The critical sequence pattern varied according to clusters. However, the differences in directiveness were discernible. The behavior of the man teacher was less directive than that of the woman teacher.

(2) Variation of sequence and variety across class time as a function of grade level

The millage matrices computed by clusters from 8 lessons of three grade levels rated by six observers (T_2), each containing 9,552 six-second-time units, are presented in Table 12. Some rows representing categories in which significant differences were found between grades levels have been identified.

In the Cluster I matrix, the lower grade level shows as dominating critical sequence patterns "silence, information, silence" (10-10, 10-5, 5-5, 5-10) and "silence, stop activity, information" (10-10, 10-4, 4-5, 5-5). The more specific critical patterns are "silence, command, silence" (10-10, 10-6, 6-6, 6-10), "pupil initiation, teacher information, pupil initiation" (9-9, 9-5, 5-5, 5-9), and "pupil initiation, teacher feedback, teacher information, pupil initiation" (9-9, 9-2, 2-5, 5-9).

In the middle grade level, the critical dominating sequence pattern was "silence, corrective information, silence" (10-10, 10-5, 5-5, 5-10). The more specific patterns were "silence, feedback, silence" (10-10, 10-2, 2-2, 2-10) and "teacher participation, information, teacher participation" (11-11, 11-5, 5-5, 5-11). Thus, there was more silent guidance, feedback, and teacher participation/ information than in the lower grade level.

In Cluster II, the same characteristics of change were identified in the analysis of pupil collective activity/passivity sequence patterns. The dominant critical sequence pattern in all grade levels was "pupil collective activity in which inter-pupil contacts and/or movement are free, range of ideas are restricted, and pupils follow instruction" (2-2, 2-5, 5-5, 5-2). Typical at the lower grade level were the sequence patterns 2-2, 2-6, 6-6, 6-2 and 6-6, 6-5, 5-5, 5-6, indicating directiveness in activity and in preparations to activity. For the middle grade level, a specific critical sequence pattern was a "totally controlled movement activity, organizing" pattern (1-1, 1-6, 6-6, 6-2), indicating directiveness in different forms. A specific critical pattern at the upper grade level was "pupils

collectively moving with free contacts, using open ideas, and following instruction" (3-3, 3-5, 5-5, 5-3).

In Cluster III, the dominating social form pattern was "divided class uniform task, complete class uniform task" (2-2, 2-1, 1-1, 1-2). Specific critical sequence patterns were formed by grade level: the lower grade level, 2-2, 2-3, 3-3, 3-2, the middle grade level, 5-5, 5-1, 1-1, 1-5, and the upper grade level, 4-4, 4-1, 1-1, 1-4. The sequence and variety of the division of labor and responsibility increased as a function of grade level.

In summary, the sequence and variety increased as a function of grade level and were related to pupil behavior. In addition, the critical sequence patterns in all clusters changed and were characterized by directiveness.

(3) Variation of sequence and variety across class time as a function of P.E. subject areas

The millage matrices computed by clusters from 6 lessons of four P.E. subject areas, rated by six observers (T_2), each containing 7,164 six-second time units, are presented in Table 13 a, b, c and d.

These graphic tables are used to illustrate the next step in which the results were analyzed by using the major PEIAC/LH-75 parameters compiled from these matrices. With the millage matrices, however, the critical sequence patterns are not discernible because there are only a limited number of time units and the information was computed from repeated measures. Therefore, the indices were also used to reduce and concentrate this information.

TABLE 11 Millage matrices for episodes by teacher

Categories

Teacher

Pupil

Teacher

Other

Activity

Passivity

Other

Situation

Teachers'talk, movement, pupils talk, other

01. Accepts, praises
02. Gives corr. feedback
03. Uses ideas dev. by pup.
04. Asks, init., term. act.
05. Presents inform., org.
06. Gives dir., comm.
07. Criticizes
08. Answers questions
09. Speaks spontan., init.
10. Silent guidance
11. Silent participation
12. Confused situation

Teacher 1 (Men, N=12)

Millage-matrix-cluster I

CAT	1	2	3	4	5	6	7	8	9	10	11	12	
1	* 1	4	0	4	11	0	0	0	1	12	1	0	
2	* 3	9	0	2	13	1	0	0	1	21	3	0	
3	* 0	0	0	0	1	0	0	0	0	1	0	0	
4	* 1	2	0	2	19	1	0	4	0	9	2	0	
5	*10	9	0	22	265	7	4	0	11	63	8	2	
6	* 0	0	0	1	8	3	1	0	0	3	0	0	
7	* 0	0	0	0	4	0	1	0	0	2	0	0	
8	* 0	0	0	0	2	0	0	0	0	0	0	0	
9	* 1	2	1	0	10	0	0	0	0	2	0	0	
10	*15	25	0	8	58	3	2	0	4	206	1	0	
11	* 1	3	0	2	7	0	0	0	0	2	38	0	
12	* 0	0	0	0	2	0	0	0	0	0	1	0	
tot	*37	58	3	45	405	20	11	5	22	327	58	5	N=14328
%		3.7	5.8	0.3	4.6	40.3	2.0	1.1	0.5	2.2	32.6	5.9	1.0

Teacher 2 (Women, N=12)

Millage-matrix-Cluster I

CAT	1	2	3	4	5	6	7	8	9	10	11	12	
1	* 1	1	0	2	11	0	0	0	0	4	0	0	
2	* 2	9	0	5	12	2	0	0	0	20	1	0	
3	* 0	0	0	0	1	0	0	0	0	0	0	0	
4	* 2	4	0	4	31	8	0	5	1	28	2	0	
5	* 6	9	0	38	279	13	1	0	5	30	2	2	
6	* 1	2	0	5	14	23	0	0	0	6	1	0	
7	* 0	0	0	0	1	0	0	0	0	0	0	0	
8	* 0	0	0	1	2	0	0	0	0	0	0	0	
9	* 0	1	0	0	4	0	0	0	0	1	0	0	
10	* 7	23	0	28	25	6	0	0	2	197	1	1	
11	* 0	1	0	1	3	1	0	0	0	0	48	0	
12	* 0	0	0	0	1	0	0	0	0	3	0	0	
tot	*23	54	3	89	391	56	4	6	11	294	58	5	N=14328
%		2.4	5.4	0.3	8.8	38.9	5.7	0.5	0.7	1.1	29.3	5.8	1.1

Pupil's collective movement activity/ passivity and social access

1. Contacts, ideas cont.
2. Contacts free, ideas cont.
3. Contacts free, ideas open
4. Pupils' spont. activity
5. Pupils follow instruction
6. Pupils organization
7. Pupils wait for turn
8. Confused situation

Millage-matrix-Cluster II

CAT		1	2	3	4	5	6	7	8	
1	*	44	0	0	0	4	0	0	0	
2	*	0	490	0	0	21	10	1	3	
3	*	0	0	74	0	2	1	0	0	
4	*	0	0	0	6	0	0	0	0	
5	*	3	24	2	0	163	14	1	1	
6	*	0	7	1	0	17	81	0	0	
7	*	0	1	0	0	0	0	1	0	
8	*	0	2	0	0	1	0	0	0	
tot	*	49	528	79	8	212	110	4	5	N=14328
%		5.0	52.6	8.0	0.8	21.1	11.0	0.4	1.1	

Millage-matrix-Cluster II

CAT		1	2	3	4	5	6	7	8	
1	*	150	1	0	0	23	2	0	0	
2	*	1	255	0	0	22	7	0	1	
3	*	0	0	77	0	2	1	0	0	
4	*	0	0	0	0	0	0	0	0	
5	*	21	22	2	0	263	18	2	2	
6	*	2	5	0	0	19	70	0	0	
7	*	1	1	0	0	1	0	3	0	
8	*	1	2	0	0	0	0	0	0	
tot	*	179	288	82	1	334	99	8	5	N=14328
%		17.8	28.8	8.2	0.1	33.3	10.0	0.8	1.0	

Social form

1. Complete class, uniform task
2. Divided class, uniform task
3. Divided class, different tasks
4. Div. cl. diff. task within gr.
5. Indivudual work, unif. tasks
6. Individual work, diff. tasks
7. Other, conf. situation

Millage-matrix-Cluster III

CAT		1	2	3	4	5	6	7	
1	*	305	2	1	0	1	0	2	
2	*	1	222	0	0	0	0	1	
3	*	1	0	199	0	0	0	0	
4	*	1	0	0	147	0	0	0	
5	*	1	0	0	0	94	0	0	
6	*	0	0	0	0	0	0	0	
7	*	2	0	0	1	0	0	5	
tot	*	313	226	202	150	97	0	10	N=14328
%		31.2	22.5	20.2	14.9	9.7	0.0	1.5	

Millage-matrix-Cluster III

CAT		1	2	3	4	5	6	7	
1	*313	2	1	0	0	0	0		
2	* 1331	2	0	0	0	1			
3	* 0	1	248	0	0	0	2		
4	* 0	0	0	28	0	0	0		
5	* 0	0	0	0	50	0	0		
6	* 0	0	0	0	0	4	0		
7	* 3	1	0	0	0	0	0		
tot	*319	336	252	30	51	5	5	N=14328	
%		31.7	33.5	25.1	3.0	5.1	0.5	1.0	

TABLE 12 Millage matrices for episodes by grade level

Categories	Lower level (N=8)													Middle level (N=8)													Upper level (N=8)												
Teachers' talk, movement, pupils talk, other	CAT	1	2	3	4	5	6	7	8	9	10	11	12	CAT	1	2	3	4	5	6	7	8	9	10	11	12	CAT	1	2	3	4	5	6	7	8	9	10	11	12	
01. Accepts, praises	1 *	1	1	0	3	11	1	0	0	1	5	0	0	1 *	0	4	0	2	10	0	0	0	0	10	0	0	1 *	1	2	0	4	12	0	0	0	0	9	1	0	
02. Gives corr. feedback	2 *	2	8	0	6	12	1	0	0	1	14	2	0	2 *	3	10	0	3	16	1	0	0	1	29	1	0	2 *	3	9	0	2	9	1	0	0	0	19	3	0	
03. Uses ideas dev. by pup.	3 *	0	0	0	0	2	0	0	0	0	1	0	0	3 *	0	0	0	0	0	0	0	0	0	0	0	0	3 *	0	0	0	0	0	0	0	0	0	0	0	0	
04. Asks, init., term. act.	4 *	1	3	1	4	23	6	1	11	1	19	1	0	4 *	3	2	0	2	24	4	0	2	1	23	2	0	4 *	0	2	0	2	28	5	0	1	0	12	4	0	
05. Presents inform., org.	5 *	9	7	0	33	310	17	4	0	14	33	5	2	5 *	7	11	0	32	244	10	3	0	6	57	3	2	5 *	8	7	0	25	262	4	0	0	3	48	7	2	
06. Gives dir., comm.	6 *	2	2	0	3	16	15	1	0	1	3	0	0	6 *	0	1	0	3	11	13	1	0	0	6	1	0	6 *	0	0	0	3	6	12	0	0	0	4	0	0	
07. Criticizes	7 *	0	0	0	1	4	0	1	0	1	2	0	0	7 *	0	0	0	0	4	1	1	0	0	2	0	0	7 *	0	0	0	0	0	0	0	0	0	0	0	0	
08. Answers questions	8 *	1	0	0	2	5	0	0	0	0	0	0	0	8 *	0	0	0	0	1	0	0	0	0	0	0	0	8 *	0	0	0	0	0	0	0	0	0	0	0	0	
09. Speaks spontan., init.	9 *	3	3	1	1	13	0	0	0	1	3	0	0	9 *	0	1	0	0	6	0	0	0	0	1	0	0	9 *	0	1	0	0	3	0	0	0	0	1	0	0	
10. Silent guidance	10*	6	18	0	18	31	4	1	0	5	168	2	0	10*	12	33	0	19	53	7	2	0	2	196	1	1	10*	14	20	0	16	40	3	0	0	1	241	0	1	
11. Silent participation	11*	0	2	0	1	5	0	0	0	0	2	17	0	11*	0	1	0	1	4	1	0	0	0	1	32	0	11*	2	3	0	3	5	0	0	0	0	2	81	0	
12. Confused situation	12*	0	0	0	0	1	0	0	0	0	2	1	0	12*	0	0	0	0	1	0	0	0	0	2	0	0	12*	0	0	0	0	1	0	0	0	0	1	1	0	
	tot*	28	50	5	76	440	47	12	13	29	258	30	5	N=9552	tot*	30	68	2	67	382	39	10	2	13	332	44	5	N=9552	tot*	32	49	1	58	372	28	0	1	634	299	5	N=9552
	%	3.0	4.5	0.6	8.5	42.3	4.4	1.3	1.1	2.7	27.3	3.3	1.0		%	3.1	6.9	0.2	6.7	38.0	3.9	1.0	0.3	1.3	33.3	4.5	1.0		%	3.1	5.0	0.1	5.9	37.2	2.9	0.0	0.2	0.7	34.0	9.9	1.0	
Pupil's collective movement activity/passivity and social access	CAT	1	2	3	4	5	6	7	8					CAT	1	2	3	4	5	6	7	8					CAT	1	2	3	4	5	6	7	8				
1. Contacts, ideas cont.	1 *	108	1	0	0	12	1	0	0					1 *	113	1	0	0	16	2	0	0					1 *	70	0	0	0	12	0	0	0				
2. Contacts free, ideas cont.	2 *	13	15	0	0	22	9	0	1					2 *	13	97	0	0	20	9	1	3					2 *	04	06	0	0	21	8	0	2				
3. Contacts free, ideas open	3 *	0	0	20	0	2	1	0	0					3 *	0	0	87	0	2	1	0	0					3 *	0	01	20	0	2	1	0	1				
4. Pupils' spont. activity	4 *	0	0	0	9	0	0	0	0					4 *	0	0	0	2	0	0	0	0					4 *	0	0	0	0	0	0	0	0				
5. Pupils follow instruction	5 *	10	25	2	02	59	21	1	2					5 *	16	24	2	01	71	13	0	1					5 *	10	20	3	02	10	14	3	1				
6. Pupils organization	6 *	1	4	0	0	25	11	5	1					6 *	2	6	1	0	16	65	0	0					6 *	0	9	0	0	14	47	0	0				
7. Pupils wait for turn	7 *	1	1	0	0	1	0	2	0					7 *	0	1	0	0	0	0	3	0					7 *	1	0	0	0	1	0	2	0				
8. Confused situation	8 *	0	3	0	0	1	0	0	0					8 *	0	1	0	0	1	0	0	0					8 *	0	2	1	0	1	0	0	0				
	tot*	123	52	24	10	325	150	6	6	N=9552				tot*	134	34	92	3	229	93	6	5	N=9552				tot*	84	39	126	0	264	72	6	5	N=9552			
	%	17.0	33.3	3.7	1.1	30.9	12.6	0.3	1.1					%	13.4	43.2	9.2	0.3	22.9	9.3	0.6	1.1					%	8.4	43.7	12.6	0.0	26.4	7.2	0.6	0.5				
Social form	CAT	1	2	3	4	5	6	7						CAT	1	2	3	4	5	6	7						CAT	1	2	3	4	5	6	7					
1. Complete class, uniform task	1 *	256	3	1	0	1	0	1						1 *	399	2	0	0	1	0	2						1 *	272	2	1	1	0	0	0					
2. Divided class, uniform task	2 *	13	68	2	0	0	0	1						2 *	12	35	0	0	0	0	1						2 *	02	26	0	0	0	0	0					
3. Divided class, different tasks	3 *	0	12	44	0	0	0	1						3 *	1	01	76	0	0	0	0						3 *	0	02	52	0	0	0	1					
4. Div. cl. diff. task within gr.	4 *	0	0	0	38	0	0	0						4 *	0	0	0	76	0	0	0						4 *	1	0	01	49	0	0	1					
5. Indivudual work, unif. tasks	5 *	1	0	0	0	51	0	0						5 *	0	0	0	0	90	0	0						5 *	1	0	0	0	75	0	0					
6. Individual work, diff. tasks	6 *	0	0	0	0	0	6	0						6 *	0	0	0	0	0	0	0						6 *	0	0	0	0	0	0	0					
7. Other, conf. situation	7 *	3	0	0	1	0	0	7						7 *	2	1	0	0	0	0	0						7 *	1	0	0	1	0	0	0					
	tot*	263	74	24	8	39	52	7	13	N=9552				tot*	406	240	17	8	77	92	0	5	N=9552				tot*	278	229	25	5	31	53	7	7	N=9552			
	%	30.3	37.5	19.7	4.9	5.2	0.6	1.8						%	40.4	23.9	17.7	7.7	9.2	0.0	1.1						%	27.7	28.8	25.4	15.3	7.8	0.0	1.1					

6 observers N=144
24 lessons (20 minutes) N=144
28 800 6 second time units

TABLE 13 Millage matrices for episodes by four subject areas of physical education

A Gymnastics (N = 6 lessons)

CAT	1	2	3	4	5	6	7	8	9	10	11	12	
1 *	1	1	0	6	16	1	0	0	1	6	2	0	
2 *	2	5	0	6	10	2	0	0	0	13	2	0	
3 *	0	0	0	0	1	0	0	0	0	0	0	0	
4 *	3	5	0	6	42	13	0	8	1	27	6	1	
5 *	9	8	0	55	259	18	4	1	7	31	9	2	
6 *	3	3	0	9	18	41	0	0	0	8	1	0	
7 *	0	0	0	1	4	0	1	0	1	0	0	0	
8 *	0	0	0	2	3	0	0	0	0	0	0	0	
9 *	1	1	0	1	8	0	0	0	0	1	0	0	
10 *	12	15	0	21	32	7	0	0	2	102	0	0	
11 *	3	3	0	5	9	1	0	0	0	1	28	0	
12 *	0	0	0	0	1	0	0	0	0	1	2	0	
tot *	40	46	3	117	409	87	10	9	17	197	55	5	N=7164
%	4.0	4.6	0.3	11.7	40.7	8.8	1.0	1.0	1.7	19.6	5.5	1.1	

CAT	1	2	3	4	5	6	7	8	
1 *	212	2	0	0	41	3	0	0	
2 *	3	267	0	0	25	8	0	1	
3 *	0	0	11	0	1	1	0	0	
4 *	0	0	0	0	0	0	0	0	
5 *	37	29	2	0	227	14	1	3	
6 *	3	4	0	0	18	58	0	0	
7 *	2	0	0	0	0	0	1	0	
8 *	1	3	0	0	0	0	0	0	
tot *	260	308	14	1	316	86	5	6	N=7164
%	26.0	30.7	1.5	0.1	31.5	8.6	0.5	1.1	

CAT	1	2	3	4	5	6	7	
1 *	575	3	0	0	0	0	1	
2 *	1	259	0	0	0	0	0	
3 *	0	0	42	0	0	0	0	
4 *	1	0	0	97	0	0	1	
5 *	0	0	0	0	0	0	0	
6 *	0	0	0	0	0	0	0	
7 *	2	0	0	1	0	0	0	
tot *	586	263	43	101	0	0	5	N=1764
%	58.2	26.3	4.4	10.0	0.0	0.0	1.0	

B Apparatus (N = 6 lessons)

CAT	1	2	3	4	5	6	7	8	9	10	11	12	
1 *	0	1	0	2	4	0	0	0	0	10	0	0	
2 *	1	2	0	2	9	0	0	0	0	20	1	0	
3 *	0	0	0	0	0	0	0	0	0	0	0	0	
4 *	1	2	0	1	22	0	0	4	0	19	0	0	
5 *	6	4	0	23	265	8	1	0	8	50	1	2	
6 *	0	0	0	0	9	1	0	0	0	2	0	0	
7 *	0	0	0	0	1	0	0	0	0	1	0	0	
8 *	0	0	0	0	2	0	0	0	0	0	0	0	
9 *	0	1	1	0	7	0	0	0	0	1	0	0	
10 *	9	23	0	22	44	3	0	0	1	339	1	2	
11 *	0	1	0	0	3	0	0	0	0	1	9	0	
12 *	0	0	0	0	2	0	0	0	0	1	0	0	
tot *	20	39	1	54	374	15	3	5	12	449	16	6	N=7164
%	4.5	10.6	0.3	3.0	45.5	2.5	0.7	0.3	2.4	27.4	1.8	1.0	

CAT	1	2	3	4	5	6	7	8	
1 *	3	0	0	0	1	0	0	0	
2 *	0	509	0	0	16	13	0	3	
3 *	0	0	15	0	0	0	0	0	
4 *	0	0	0	11	0	0	0	0	
5 *	0	20	0	0	251	16	0	1	
6 *	0	11	0	0	18	92	0	0	
7 *	0	0	0	0	0	0	0	0	
8 *	0	0	0	0	2	0	0	0	
tot *	5	542	16	12	290	123	2	5	N=7164
%	0.5	54.0	1.6	1.3	29.0	12.3	0.2	1.1	

CAT	1	2	3	4	5	6	7	8	9	10	11	12	
1 *	0	0	0	1	8	0	0	0	0	4	0	0	
2 *	1	5	0	2	8	0	0	0	0	10	3	0	
3 *	0	0	0	0	1	0	0	0	0	2	0	0	
4 *	1	2	1	3	22	3	1	5	1	22	2	0	
5 *	3	3	0	30	244	6	2	0	6	46	5	2	
6 *	0	0	0	2	8	7	1	0	0	2	1	0	
7 *	0	0	0	0	1	4	0	0	0	2	0	0	
8 *	0	0	0	0	1	1	0	0	0	0	0	0	
9 *	0	0	0	1	1	5	0	0	0	1	1	0	0
10 *	8	15	0	20	40	4	2	0	2	222	3	0	
11 *	0	3	0	0	2	5	0	0	0	0	2	126	0
12 *	0	0	0	0	0	0	0	0	0	3	0	0	
tot *	17	32	5	68	352	25	10	5	12	321	143	5	N=7164
%	1.8	3.3	0.5	6.8	35.0	2.5	1.0	0.5	1.3	32.0	14.3	1.0	

CAT	1	2	3	4	5	6	7	8	
1 *	30	0	0	0	4	1	0	0	
2 *	0	287	0	0	27	6	2	0	
3 *	0	0	276	0	7	3	0	1	
4 *	0	0	0	3	0	0	0	0	
5 *	3	26	8	0	177	14	4	1	
6 *	0	4	2	0	17	58	0	0	
7 *	1	2	0	0	2	0	6	0	
8 *	0	2	1	0	0	0	0	0	
tot *	36	325	291	4	237	85	14	5	N=7164
%	3.6	32.4	29.0	0.5	23.6	8.5	1.4	1.0	

CAT	1	2	3	4	5	6	7	
1 *	329	3	1	0	4	0	1	
2 *	2	144	1	0	1	0	0	
3 *	0	1	142	0	0	0	0	
4 *	0	0	0	52	0	0	0	
5 *	4	1	0	0	285	0	1	
6 *	0	0	0	0	0	9	0	
7 *	4	0	0	0	0	0	0	
tot *	341	150	146	53	292	10	5	N=1764
%	34.0	15.0	14.6	5.3	29.1	1.0	1.0	

C Rhythmic Movement Expression (N = 6 lessons)

D Ball Games (N = 6 lessons)

CAT	1	2	3	4	5	6	7	8	9	10	11	12	
1 *	0	1	0	2	4	0	0	0	0	10	0	0	
2 *	1	2	0	2	9	0	0	0	0	20	1	0	
3 *	0	0	0	0	0	0	0	0	0	0	0	0	
4 *	1	2	0	1	22	0	0	4	0	19	0	0	
5 *	6	4	0	23	265	8	1	0	8	50	1	2	
6 *	0	0	0	0	9	1	0	0	0	2	0	0	
7 *	0	0	0	0	1	0	0	0	0	1	0	0	
8 *	0	0	0	0	2	0	0	0	0	0	0	0	
9 *	0	1	1	0	7	0	0	0	0	1	0	0	
10*	9	23	0	22	44	3	0	0	1	339	1	2	
11*	0	1	0	0	3	0	0	0	0	1	9	0	
12*	0	0	0	0	2	0	0	0	0	1	0	0	
tot*	20	39	1	54	374	15	3	5	12	449	16	6	N=7164
%	2.0	3.9	0.2	5.4	37.4	1.6	0.4	0.5	1.3	44.7	1.7	1.1	

CAT	1	2	3	4	5	6	7	8	
1 *	143	1	0	0	8	1	0	0	
2 *	1	429	0	0	16	7	0	2	
3 *	0	0	1	0	0	0	0	0	
4 *	0	0	0	0	0	0	0	0	
5 *	8	17	0	0	199	20	0	1	
6 *	1	5	0	0	21	94	0	0	
7 *	0	0	0	0	1	0	2	0	
8 *	0	3	0	0	1	0	0	0	
tot*	156	457	2	0	248	124	4	5	N=7164
%	15.5	45.5	0.2	0.1	24.7	12.4	0.5	1.1	

CAT	1	2	3	4	5	6	7	
1 *	59	1	1	0	0	0	0	
2 *	0	669	0	1	0	0	3	
3 *	0	0	80	0	0	0	0	
4 *	1	0	0	168	0	0	0	
5 *	0	0	0	0	1	0	0	
6 *	0	0	0	0	0	0	0	
7 *	0	3	0	0	0	0	0	
tot*	63	676	82	170	1	0	5	N=1764
%	6.3	67.3	8.2	17.0	0.2	0.0	1.0	

Summary

In each of the sequence patterns presented and discussed so far, decisions were required of the teacher for critical transitions, that is, sequence pairs with different numbers. In steady state cells sequence pairs have the same number.

The sequence and variety in the three cluster matrices were different, as expected. In Cluster I, more than one half of all sequence pairs were in steady state cells, in Cluster II more than 80%, and in Cluster III more than 90%. The critical decisions concerning social form, division of labor and responsibility, and the forms of pupil collective activity/passivity were the general dominating aspects when teaching behavior was analyzed.

Variation according to context variables, gender, grade level and subject area of P.E. was analyzed and described. The male and female teachers were quite homogeneous but flexible. The sequence and variety were related to different aspects. However, the male teacher was in general less directive. Changes in "critical" teaching behavior appeared as functions of the grade level. The directiveness of the teacher decreased as the age of pupils increased. At the same time, the teacher's silent guidance, participation, use of pupils' ideas, and pupils' responsibility increased, as did the variety of critical sequence patterns.

The interpretation and comparison of matrices describing the instructional process as function of subject areas of physical education are made in the next step. The results of the major PEIAC/LH-75 parameters, computed from the primary and secondary information of these matrices, are presented and discussed. The displays presented in the four parts of Table 13 a, b, c and d are used to enhance and clarify this description.

6.2.4 Describing the instructional process with the major PEIAC/LH-75 parameters and indices according to contextual variation: gender (1), grade level (2) and P.E. subject area (3)

Further analysis included a comparison of the means of each interaction process across class time with PEIAC/LH-75 parameters (Table 3, p. 87). The indices are based on unit coding and the statistical procedures are based on category frequencies, percents and ratios. These are computed separately from matrices of the three clusters. The significance of the differences between the means of PEIAC/LH-75 indices variation according to context variables (gender grade level and P.E. subject areas) was estimated by using the Mann-Whitney U-test and the rank order was determined by functions of variability. The results are presented in Tables 14, 15 and 16, and the statistical differences of the means of PEIAC/LH-75 indices by frame factor are summarized in Table 17.

The indices were used to reduce the data and to give a concentrated picture of the elements of this category system grouped into three clusters.

(1) Variation of the means of PEIAC/LH-75 indices according to context variables variation according to context: gender

The significance of differences between indices (presented in table 3, p 87) estimated for the 12 lessons of the male and female teachers, rated by six coders (T_2), and containing 14,328 six-second time units, are presented in Table 14.

The differences of male and female teachers' initiation/response behavior were reflected in pupil behavior. The "pupil verbal initiation ratio" (PVIR), "nonverbal initiation ratio" (PIR), and percent of "pupil collective movement activity" were higher for the male teacher than for the female teacher. The differences in the means of these indices were statistically significant. On the other hand, the "teacher question and activity initiation/termination ratio" (TQAR) and the "pupil collective following instruction, organizing ratio" (PIOR) were higher for the female teacher. The differences in the means of these indices were also statistically significant. The "teacher response ratio" (TRR), based on verbal behavior, was only slightly higher for the man teacher than for the woman teacher.

TABLE 14 Significance of differences between PEIAC/LH-75 indices estimated for two teachers (man-woman) (T_2), Mann-Whitney U-test

No	Symbol	Name of Index	Teacher 1. (n=12 h)	Rank	Teacher 2. (n=12 h)	Rank	Differences: Mann-Whitney U-test 1.-2. z
1	TT	Percent teacher talk	58.17	2.	62.35	1.	-0.98
2	PT	Percent pupil talk	2.73	1.	1.79	2.	-1.36
3	TSAR	Teacher sustained activity ratio	53.04	2.	56.62	1	-0.64
4	TSGPR	Teacher silent guidance and participation ratio	39.88	1.	36.13	2.	-0.58
5	TRR	Teacher response ratio	30.57	1.	28.14	2.	-0.17
6	TQAR	Teacher question and activity initiation-termination ratio	10.13	2.	18.63	1.	-2.37**
7	CCR	Content emphasis ratio	45.11	2	48.07	1.	-0.35
8	PVIR	Pupil verbal initiation ratio	81.33	1.	62.65	2.	-3.33***
9	PIR	Pupil initiation ratio (verbal and nonverbal)	94.57	1.	77.85	2.	-1.85*
10	TPR	Teacher praise ratio	77.22	2.	84.07	1.	-0.61
11	PCA	Percent pupil collective activity	66.69	1.	55.18	2.	-2.54**
12	PSUAR	Pupil sustained activity ratio	86.48	1.	82.22	2.	-1.56
13	PSAR	Pupil social access ratio	13.24	2.	15.20	1.	-0.29
14	PIOR	Pupil collective following instruction, organizing ratio	32.27	2.	43.43	1.	-2.48**
15	SGWR	Pupil social group work ratio	35.65	1.	6.26	2.	-0.46
16	PIWR	Pupil individual work ratio	9.81	1.	5.71	2.	-0.34
17	SFVR	Social form variability ratio	7.00	1.	7.00	2.	-0.06
18	SSFR	Sustained social form ratio	97.38	2.	97.65	1.	-0.96

6 observers
n = 14328 6 second time units

Levels of significance
* =p < 0.05
** = p < 0.01
*** = p < 0.001

(2) Variation of the means of PEIAC/LH-75 indices according to context: grade level

The significance of differences between indices as estimated for the 8 lessons of three grade levels, rated by six coders (T_2), and containing 9,552 six-second time units, are presented in Table 15, p. 116).

The differences between the instructional processes of the three grade levels were clearly recognized in the parameters describing the general features of the use of time, such as the indices describing pupil verbal/nonverbal behavior and pupil collective movement activity/ passivity. The percent of class time used for "pupil talk" (PT) decreased at the middle and upper grade levels (from 4% to 0.86%), whereas the amount of "teacher's silent guidance and participation (TSPGR) increased (from 23% to 44%). At the middle and upper grade levels, the "teacher verbal praise ratio" (TPR) increased. The differences in these indices were statistically significant. The percent of "pupil collective activity" (PCA) increased at the middle grade level (51% to 66%), whereas the ratio describing pupil collective passivity, in which the "pupils follow instruction organize themselves" (PIOR), decreased (47% to 32 %). These differences between indices were statistically significant, the "pupil individual work ratio (PIWR) was at its highest at the middle gread level, but differences in this variable between grade levels were not statistically significant (Table 15).

(3) Variation of the means of PEIAC/LH-75 indices according to context: P.E. subject areas

The significance of differences between indices as estimated for 6 lessons of four P.E. subject area, rated by six coders (T_2), and containing 7,164 six-second time units, are presented in Table 16. Differences in indices were strongly related to the content of the subject areas. Fourteen of the eighteen indices produced statistically significant differences. These will be presented by referring to the rank order of the indices.

The percent of class time devoted to "teacher talk" (TT) was highest (71%) in gymnastics and lowest in rhythmic movement expression (50%) and ball games (51%). Both the "teacher sustained activity ratio" (TSAR) and the "teacher silent guidance and participation ratio" (TSGPR) were highest in ball games and rhythmic movement expression and lowest in gymnastics. The "teacher response ratio" (TRR), which was adapted from Flanders' ID-ratio, was highest in gymnastics, second highest in apparatus and lowest in ball games.

Typical of gymnastics was a high percentage for the "teacher question, activity initiation and termination ratio" (TQAR). This ratio was second highest in rhythmic movement expression and lowest in apparatus. The "content emphasis ratio" (CCR) was highest in gymnastics and second highest in apparatus and lowest in rhythmic movement expression.

TABLE 15 Significance of differences between PEIAC/LH-75 indices estimated for three grade levels (T_2), Mann-Whitney U-test

No	Symbol	Name of Index	Lower level 1. (N=8 h)	Rank	Middle level 2. (N=8 h)	Rank	Upper level 3. (N=8 h)	Rank	Mann-Whitney U-test Differences: 1.-2. z	1.-3. z	2.-3. z
1	TT	Percent teacher talk	66.19	1.	60.13	2.	54.46	3.	-1.26	-1.58	-0.42
2	PT	Percent pupil talk	4.28	1.	1.64	2.	0.86	3.	-2.63**	-3.05***	-1.68*
3	TSAR	Teacher sustained activity ratio	53.06	2.	50.23	3.	61.20	1.	-0.42	-1.37	-1.79*
4	TSGPR	Teacher silent guidance and participation ratio	23.90	3.	40.96	2.	44.78	1.	-1.68*	-1.79*	-0.74
5	TRR	Teacher response ratio	26.64	3.	27.74	2.	33.09	1.	-0.11	-1.47	-0.74
6	TQAR	Teacher question and activity initiation-termination ratio	14.88	2.	14.94	1.	13.64	3.	-0.53	-0.53	-0.21
7	CCR	Content emphasis ratio	51.71	1.	44.92	2.	43.13	3.	-1.37	-1.58	-0.42
8	PVIR	Pupil verbal initiation ratio	69.44	2.	17.83	3.	80.49	1.	-0.79	-0.74	-0.32
9	PIR	Pupil initiation ratio (verbal and nonverbal)	76.43	2.	32.25	3.	99.95	1.	-1.16	-1.26	-0.11
10	TPR	Teacher praise ratio	48.25	3.	74.87	2.	98.73	1.	-1.47	-2.80**	-2.70**
11	PCA	Percent pupil collective activity	51.24	3.	66.49	1.	65.06	2.	-2.31**	-1.89*	-0.37
12	PUAR	Pupil sustained activity ratio	83.11	3.	84.10	2.	85.85	1.	-0.42	-0.84	-0.21
13	PSAR	Pupil social access ratio	6.99	3.	14.42	2.	19.46	1.	-0.74	-1.00	-0.54
14	PIOR	Pupil collective following instruction, organizing ratio	47.51	1.	32.30	3.	33.76	2.	-2.31**	-1.89*	-0.32
15	SGWR	Pupil social group work ratio	29.20	3.	38.51	2.	41.07	1.	-0.53	-0.63	-1.47
16	PIWR	Pupil individual work ratio	6.10	3.	9.27	1.	7.84	2.	-0.72	-0.72	-0.14
17	SFVR	Social form variability ratio	7.00	1.	7.00	1.	6.00	3.	-0.49	-0.50	-0.00
18	SSFR	Sustained social form ratio	97.25	3.	97.69	1.	97.60	2.	-1.27	-0.32	-0.32

6 observers
N = 9552 6 second time units

Levels of significance
* = $p<0.05$
** = $p < 0.01$
*** = $p < 0.001$

TABLE 16 Significance of differences between PEIAC/LH-75 indices estimated for four subject areas (T_2), Mann-Whitney U-test

No	Symbol	Name of Index	Gymnastics 1. (N=6 h)	Rank	Apparatus 2. (N=6 h)	Ra.	Rhythmic movement express 3. (N=6 h)	Ra.	Ball games 4. (N=6 h)	Ra.	Mann-Whitney U-test Differences: 1.-2. z	1.-3. z	1.-4. z	2.-3 z	2.-4. z	3.-4. z
1	TT	Percent teacher talk	71.48	1	67.46	2	51.13	3	50.98	4	-0.80	-2.56^{xx}	-2.72^{xx}	-2.08^{x}	-2.56^{xx}	0.00
2	PT	Percent pupil talk	2.73	1	2.71	2	1.82	3	1.77	4	-0.40	0.00	-0.16	-0.96	-1.28	-0.08
3	TSAR	Teacher sustained activity ratio	44.81	4	50.95	3	61.18	2	62.38	1	-1.76^{x}	-2.40^{xx}	-2.56^{xx}	-1.44	-1.92^{x}	-0.56
4	TSGPR	Teacher silent guidance and participation ratio	26.11	4	30.29	3	47.64	2	47.77	1	-0.80	-2.24^{xx}	-2.40^{xx}	-2.24^{xx}	-2.56^{x}	-0.32
5	TRR	Teacher response ratio	38.25	1	35.99	2	35.79	3	19.28	4	-0.32	-0.16	-2.08^{x}	0.00	-2.56^{xx}	-1.76^{x}
6	TQAR	Teacher question and activity initiation-termination ratio	22.37	1	6.10	4	16.27	2	12.70	3	-2.88^{xx}	-1.28	-1.44	-2.24^{xx}	-1.28	-0.80
7	CCR	Content emphasis ratio	52.74	1	48.71	2	42.04	4	42.86	3	-1.44	-1.92^{x}	-2.40^{xx}	-1.12	-0.80	-0.32
8	PVIR	Pupil verbal initiation ratio	73.78	2	88.14	1	70.23	3	61.65	4	-1.36	-1.13	-0.16	-0.32	-1.92^{x}	-0.97
9	PIR	Pupil initiation ratio (verbal and nonverbal)	66.25	4	93.24	2	115.25	1	72.08	3	-1.76^{x}	-2.08^{x}	-0.48	-1.44	-2.42^{xx}	-2.24^{xx}
10	TPR	Teacher praise ratio	79.72	3	86.13	1	10.93	4	84.39	2	-0.16	-0.49	-0.65	-0.48	-0.56	-0.16
11	PCA	Percent pupil collective activity	58.58	3	57.73	4	65.79	1	61.63	2	-0.00	-0.96	0.00	-0.96	-0.32	-0.48
12	PUAR	Pupil sustained activity ratio	77.85	4	88.47	1	83.98	3	87.12	2	-2.88^{xx}	-1.92^{x}	-2.56^{xx}	-1.44	-0.48	-0.96
13	PSAR	Pupil social access ratio	2.74	3	5.10	2	45.02	1	0.43	4	-0.33	-1.93^{x}	-0.82	-2.26^{xx}	-0.08	-2.41^{xx}
14	PIOR	Pupil collective following instruction, organizing ratio	40.30	2	41.49	1	32.28	4	37.35	3	0.00	-1.44	-0.00	-0.96	-0.32	-0.80
15	SGWR	Pupil social group work ratio	14.56	4	68.30	1	20.10	3	25.48	2	-2.89^{xx}	-0.32	-0.32	-2.72^{xx}	-1.92^{x}	-0.00
16	PIWR	Pupil individual work ratio	0.00	4	0.23	2	30.47	1	0.20	3	-1.00	-3.08^{xxx}	-1.00	-2.99^{xxx}	-0.12	-2.99^{xxx}
17	SFVR	Social form variability ratio	5.00	4	6.00	2	7.00	1	6.00	2	-0.70	-2.00^{x}	-1.55	-1.24	-0.58	-1.08
18	SSFR	Sustained social form ratio	97.96	2	97.77	3	96.33	4	97.99	1	-0.08	-2.72^{xx}	-0.08	-2.40^{xx}	-0.32	-2.72^{xx}

N=7164 6 second time units
6 observers

Levels of significance
x = p< 0.05
xx = p < 0.01
xxx = p < 0.001

The "pupil verbal initiation ratio" (PVIR) was highest in apparatus and lowest in ball games. The variability of "pupil verbal and nonverbal initiation ratio" (PIR) was great. It was highest in rhythmic movement expression and lowest in gymnastics and ball games. The "pupil sustained movement activity ratio" (PSAR) was highest in apparatus, second highest in ball games and lowest in gymnastics.

The "pupil social access ratio" (PSAR), measured with "pupil movement activity", was strongly related to the subject area. It was highest in rhythmic movement expression and lowest in ball games. It should be noted also that the differences between the indices describing the division of labor and responsibility were clearly related to the content of the subject area.

The "pupil social group work ratio" (SGWR) was highest in apparatus and lowest in gymnastics, whereas the "individual work ratio" (PIWR) was high only in rhythmic movement expression and could not be estimated for gymnastics with this data. The "sustained social form ratio" (SSFR) was highest in ball games and lowest in rhythmic movement expression.

In only four of the eighteen indices were the differences between subject areas not statistically significant. These were such general characteristics as "pupil talk" (PT), "pupil collective activity ratio" (PCA), and "pupil collective following instruction, organizing ratio" (PIOR). As stated earlier, these characteristics are all related to pupil behavior, and thus to grade level.

6.2.5 Summary

In the variation of all 18 main parameters of the PEIAC/LH-75 system according to context variables, gender, grade level and P.E. subject area statistically significant differences were found. Five of these differences were related to gender, six to grade level, and fourteen to the subject areas in physical education (Table 17).

The teaching behavior of the male and the female teacher in this study was quite homogeneous in many different contexts and they were rather flexible in their behavior. However, a difference between the teachers' initiation response behavior was discernible. It was obviously related to pupil behavior (age) and appeared also to reflect the training background of the teachers. Within the teacher response behavior parameter, the praise ratio increased in upper grade level and content-centeredness diminished mainly in response to pupil behavior (grade level effect).

The influence of subject specific content on the instructional process was dominant and was reflected in the different aspects indicative of initiation response behavior. The main point was thus, not only the subject matter of physical education as such, but the kinds of content it consisted of, and how the instructional processes were arranged to accommodate them. The temporal basis of the instructional process, described, e.g., by analyzing the amount of teacher talk, silent guidance and participation, as well as teacher sustained activity, pupil sustained activity and sustained social form ratios, was clearly related to contextual variables in terms of the content of the subject area.

The social forms, division of labor and responsibility between teacher and pupils, and among pupils, were strongly related to the content and quality of the subject matter. Thus, the pre-interactive decisions related to context variables strongly determined the variation of the instructional process and its progress across time.

It can be concluded that the major PEIAC/LH-75 parameters were able to provide concentrated information about the directiveness/non-directiveness of teacher behavior and about how the frame factors used in this study of the teaching process in P.E. classes. The importance of preserving the sequence when categorizing these three aspects of teaching was emphasized in the results of this study.

6.3 Phase II: Reliability and objectivity of coding

This section of the study results deals with the problem of the reliability of coding attached to the use of the observation system PEIAC/LH-75. In research work using observational systems, the testing of hypotheses requires that the observation system employed possesses sufficiently high reliability. Therefore, in developing and constructing a measuring instrument, it is crucial to provide data pertaining to reliability, as well as to discuss which reliability measures were selected and why. The question of the reliability of observation systems is a complicated one because the classification system and coder together constitute the measuring instrument. Therefore, in evaluating its usefulness, attention must be paid both to the quality of the information utilized and to the way in which it is used in the coding process. Because the value of results in observational studies depends crucially on the manner in which the instrument has been used in the coding process, an effort is made in the present study to concentrate on these aspects of evidence associated with reliability, that is, on the objectivity of coding. In this context, it signifies the degree of independence between the final results of coding and the coder himself (Komulainen 1970, 1973).

In examining the overall reliability of this observation instrument, the customary profile method, or total-events-approach, of Scott (1955) was applied. It was also considered appropriate to apply a method used in non-parametric measurement, the coefficient of concordance (W) elaborated by Kendall, to examine the reliability of various individual categories and to determine the applicability of various methods in examining the problem of objectivity of coding. Because this is a multidimensional classification system, every dimension had to be studied both separately and in conjunction with other clusters.

TABLE 17 Summary of the significance of differences between PEIAC/LH-75 indices estimated for two teachers, three grade levels and four subject areas (T_2), Mann-Whitney U-test

			Teachers N=12	Grade levels N=6			Subject areas N=6					
			Man-Woman	L-M	L-U	M-U	G-A	G-R	G-B	A-R	A-B	R-B
No	Symbol	Name of Index	z	z	z	z	z	z	z	z	z	z
1	TT	Percent teacher talk						xx	xx	x	xx	
2	PT	Percent pupil talk		xx	xxx	x						
3	TSAR	Teacher sustained activity ratio				x	x	xx	xx		x	
4	TSGPR	Teacher silent guidance and participation ratio		x	x			xx	xx	xx	xx	
5	TRR	Teacher response ratio							x		xx	x
6	TQAR	Teacher question and activity initiation-termination ratio	xx				xx			xx		
7	CCR	Content emphasis ratio						x	xx			
8	PVIR	Pupil verbal initiation ratio	xxx								x	
9	PIR	Pupil initiation ratio (verbal and nonverbal)	x				x	x			xx	xx
10	TPR	Teacher praise ratio			xx	xx						
11	PCA	Percent pupil collective activity	xx	xx	x							
12	PUAR	Pupil sustained activity ratio					xx	x	xx			
13	PSAR	Pupil social access ratio						x		xx		xx
14	PIOR	Pupil collective following instruction, organizing ratio	xx	xx	x							
15	SGWR	Pupil social group work ratio					xx			xx	x	
16	PIWR	Pupil individual work ratio						xxx		xxx		xxx
17	SFVR	Social form variability ratio						x				
18	SSFR	Sustained social form ratio						xx		xx		xx

Levels of significance
x = $p < 0.05$
xx = $p < 0.01$
xxx = $p < 0.001$

L = Lower level
M = Middle level
U = Upper level

A = Apparatus
B = Ball games
G = Gymnastic
R = Rhythmic movement express

The purpose of this phase of the study, then, was (1) to determine (a) the within-occasion reliability (agreement) and (b) between-occasion reliability (constancy) (i) by cluster, (ii) by coder pair, (iii) by situation, and (iv) by content of lessons, (2) to examine the reliability of the various individual categories (a) by category and by cluster, (b) between clusters, and (c) by coding occasion, and (3) to examine the applicability of the different methods used for assessing the reliability of a multidimensional observation instrument.

6.3.1 Results concerning overall reliability

The reliability components, within-occasion reliability (agreement) and between-occasion reliability (constancy), were examined by clusters, by coder pairs, by different coding circumstances and by different content situations of physical education classes (teacher, grade level and subject area). The final results give some idea of the experimental use of the observation instrument and of the variation in the level of mean values for different reliability components in the three clusters.

In evaluating the results it must be remembered that the number of categories in the three clusters is not equal, but 12, 8 and 7, respectively. The estimated role of chance, which is subtracted in Scott's *pi*, decreases as the number of categories increases. Thus, the probable role of chance was the least in the Verbal Cluster I. The relative frequency of occurrence of the categories is also taken into consideration by using Scott's coefficient. The mean values were highly sensitive to extreme variations and the range of variation of the six coders' coefficients by pair was shown to be remarkably wide.

A total of 8424 Scott's coefficients were computed. The differences of the means of coefficients were examined with the use of a t-test, and in some cases with both a t-test and a one-way analysis of variance (ANOVA). This method was chosen because the groups to be compared were usually more than two in number. A total of 1252 t-tests and 63 ANOVAs were computed. In interpreting the t-test, the effect of overlapping classifications at the risk-level limit was taken into consideration and thus the chosen risk-level of t values for $p>.01$ was not regarded as significant (not underlined in the tables).

Clusters (I, II, III)

The differences between clusters are presented in Table 18. The average level of mean coefficient values by cluster was rather low (.61, .65 and .69), and varied greatly between the different reliability components. An inter-coder agreement of .65, a within-coder constancy of .69 and a between-coder constancy of .60 were indicated in the scores of the video-recorded observations.

TABLE 18 Analysis by cluster: Inter-coder agreement, within-coder constancy and between-coder constancy. Mean values and standard deviations of Scott's Pi coefficients by cluster (I, II, III) and by occasion (T_1, T_2, T_3)

	Cluster I (Verbal)		Cluster II (Movement & Social Access)		Cluster III (Social Form)	
	X	SD	X	SD	X	SD
Inter-coder Agreement (n=360)						
Live Situation (T_1)	.57	.17	.61	.26	.75	.28
Videotape Recording 1 (T_2)	.61	.18	.71	.22	.77	.36
Videotape Recording 2 (T_3)	.61	.19	.59	.36	.60	.59
Within-coder Constancy (n=144)						
T_1-T_2	.66	.15	.59	.28	.62	.48
T_2-T_3	.71	.13	.66	.31	.69	.47
Between-coder Constancy (n=720)						
T_1-T_2	.54	.18	.56	.30	.61	.47
T_2-T_3	.59	.19	.62	.32	.62	.54

Examining the mean Scott's *pi* coefficient values of the coding system and the corresponding standard deviations for the videotaped observations, systematic differences in *inter-coder agreement* between clusters may be noted (Table 19). The mean coefficient values of Cluster I (Verbal) were the lowest and their standard deviations the smallest. There was no difference between the mean coefficient values in the videotaped material coding occasions T_2 and T_3. In Cluster II (Movement and Social Access), the mean values were slightly higher than those in the previous cluster and the range of standard deviations was larger. A great mean value variation (.71-.59) and statistically significant difference was noted in this cluster between the two videotape coding occasions T_2 and T_3. In Cluster III (Social Form), the mean coefficient values were the highest and the range of standard deviations the greatest.

TABLE 19 Analysis by cluster: Differences in means of Scott's Pi coefficients computed separately by cluster (I, II, III) and by occasion (T_2 and T_3) (P=<.01)

	Cluster I (Verbal)		Cluster II (Movement & Social Access)		Cluster III (Social Form)		Differences I-II	I-III	II-III
	X	SD	X	SD	X	SD	t	t	t
Inter-coder Agreement									
T_2	.61	.18	.71	.22	.77	.36	-6.67	-7.53	-2.70
T_3	.61	.19	.59	.36	.60	.59	0.93	0.31	-0.27
n=360, df=718									
Within-coder Constancy									
T_2-T_3	.71	.13	.66	.31	.69	.47	1.78	0.49	-0.64
n=144, df=286									
Between-coder Constancy									
T_2-T_3	.59	.19	.62	.32	.62	.54	-2.16	-1.41	0.00
n=720, df=1438									

Differences in inter-coder agreement between all clusters were found to be statistically significant in the first videorecorded observation (T_2), but in the second videotape recording no statistically significant differences were found. The main difference between clusters was thus the constancy of variation in the

inter-coder agreement coefficient level between coding occasions. This variation was smallest in Cluster I and greatest in Cluster II.

The comparison of the mean Scott's *pi* coefficient values showed that the values for *within-coder constancy* were higher than for inter-coder agreement and between-coder constancy in all clusters. The differences were quite noticeable in the Verbal Cluster I, where the level of the reliability coefficients as a whole was highest (.71). Also, the mean standard deviations of the Scott's *pi* coefficients varied noticeably between clusters (.13, .31, .47). However, statistically significant differences were not found between the mean coefficient values in the different clusters.

The level of *between-coder constancy* coefficients was found to be lower than the other reliability components in all clusters, and was the lowest (.59) in the verbal cluster. The differences between clusters were not found to be statistically significant.

In view of the results, it can be stated that the coding of the verbal cluster deviated from the other two clusters, among other things, in the systematic character of the between-coder constancy variation. The observers' coding of verbal events was more constant, but the differentiation between coders increased. Because this differentiation was not, however, reflected in a decreasing level of inter-coder agreement (T_3), it was apparent that the differences between coders were somehow compensated for in this cluster. In the other clusters the differences in between-coder constancy coefficients were minor, and differentiation was reflected in the decreasing level of inter-coder agreement (T_3). This differentiation of coders was, however, fairly random in character.

The structure of the coding system, the coders' behavior, and the characteristics of the coded phenomena were reflected in the results. The observation of verbal, logical communication was apparently more familiar to the coders and the interpretation of its features more constant than the observation of the other features of communication (non-verbal). The quality of the target of observation, such as tempo variation, was reflected in the results. The possible coding differences were more outstanding when a slowly changing phenomenon, such as the social form, was in question. This was found to be the case, for instance, in the considerable variation of the mean value levels within clusters.

Comparisons of live observation and videotape recording

A comparison of the *inter-coder agreement* coefficients by occasion shows them to be the lowest in the coding of the live situation (T_1) in all clusters when compared with the first videotaped observation (T_2) (Table 20). A statistically significant difference was found between the lowest cluster (Cluster I) and the other clusters. Statistically significant differences were also in evidence between the live and the videotaped observations in Clusters I and II. The greatest change occurred in the coding of Cluster II. This may be partially due to the fact that the TV screen reduces and limits the perspective of these activities for all observers and, consequently, the inter-coder agreement was increased.

Although the voices were also reduced in the recorded material, the recording may have had a more detrimental effect on visibility than on audibility.

TABLE 20 Analysis by occasion: inter-coder agreement. Significance of differences in means of Scott's Pi coefficients by cluster (I, II, III) and by occasion (T_1, T_2, T_3) (n=360, df=718, p<.01)

Cluster	T_1		T_2		T_3		Differences		
							T_1-T_2	T_1-T_3	T_2-T_3
	X	SD	X	SD	X	SD	t	t	t
Cluster I (Verbal)	.57	.17	.61	.18	.61	.19	-3.06	-2.97	-0.00
Cluster II (Movement & Social Access)	.61	.26	.71	.22	.59	.36	-5.56	0.85	5.39
Cluster III (Social Form)	.75	.28	.77	.36	.60	.59	-0.83	4.35	4.66

When comparing the Mean Scott's *pi* coefficient values of *within-coder constancy* (Table 21) observed in the live situation and from the videotaped material (T_1-T_2) with the within-coder constancy coefficient mean values of the videotape recordings (T_2-T_3) it was noted that the latter constancy coefficients were higher in all clusters. This difference between the mean coefficient values was also found to be statistically significant in Cluster I. The within-cluster variation in the level of mean coefficient values was in accordance with the previous findings in that the coefficients were highest in Cluster I and lowest in Cluster II. Also the variation of standard deviations between clusters was found to be similar to the within-coder constancy variation in general (T_2-T_3) (.15, .28, .48). Obviously the same factors which influenced cluster variation in within-coder constancy (see Table 21) also influenced variation in between-situation constancy. However, the low level of the reliability coefficients in Cluster II is indicative of the fact that the observers coded the live situation differently than the videotaped one in which some of the 'live' elements were missing due to the nature of the recording. Apparently, the two data collecting methods, direct observation and coding of recorded material, did not always produce the same observations.

As before, it appeared that in different coding situations (T_1-T_2) *between-coder constancy* coefficients were lower than the other reliability coefficient values in all clusters. The variation of mean values between the clusters was noticeable (.54, .56, .61) and similar to the general character of between-coder constancy variation (T_1-T_2). Statistically significant differences were found in the mean Scott's *pi* coefficient values (T_2-T_3) between Cluster I and Cluster II.

TABLE 21 Analysis by occasion: coder constancy. Significance of differences in means of Scott's Pi coefficients by cluster (I, II, III) and by occasion (T_1-T_2, T_2-T_3) ($p<.01$)

	T_1-T_2		T_2-T_3		Differences T_1-T_2 and T_2-T_3
	X	SD	X	SD	t
Within-coder Constancy					
Cluster I	.66	.15	.71	.13	-3.01
Cluster II	.59	.28	.66	.31	-2.00
Cluster III	.62	.48	.69	.47	-1.25
(n=144, df=286)					
Between-coder					
Cluster I	.54	.18	.59	.19	-5.12
Cluster II	.56	.30	.62	.32	-3.67
Cluster III	.61	.47	.62	.54	-0.38
(n=720, df=1438)					

An examination of the results indicates that, in spite of the circumstance variation, roughly the same general character of reliability coefficient variation was found within all the three clusters as well as between the clusters. This variation appeared to be most systematic in the coding results of Cluster I, and a result of the structure of the coding system, the observer's way of using it and the quality of the coding target. However, there is reason to assume that the coding situation partially influenced the low level of between-coder constancy coefficients in Cluster II (.56). It was apparent that the observers, when coding the videotaped material, were in fact observing a changed situation in which the 'live' elements were partially obliterated. Thus, the coding was carried out in greater agreement than in the live situation.

Coding content constancy

Coding content constancy was defined as the independence of the final results from the constancy of the coding target in different reliability components: inter-coder agreement, within-coder constancy and between-coder constancy.

In this study, the constancy variation was examined for the coding targets of two teachers, three grade levels and four physical education subject areas. The six coders' mean values and standard deviations are presented in the following tables by cluster and by reliability components with the results of the statistical significance test of the differences between the content mean values.

An overview of these results and their comparison with the previously presented general results show that the consistency of the observed phenomenon might have some systematic influence on the variation of the reliability component level in different clusters.

Teacher: When the lessons of two different teachers were the target of observation, the reliability coefficients differed systematically by reliability component and by cluster.

Inter-coder agreement varied from teacher to teacher in all clusters and in all coding occasions. The inter-coder agreement coefficient level varied according to the teacher so that in Cluster III the male teacher's coefficients were higher,

but in Cluster I and in Cluster II the situation was reversed. The mean coefficient differences were found to be statistically significant (Table 22).

For *within-coder constancy*, in the coding of the videotaped material (T_2-T_3), no statistically significant differences were found between teachers (Table 29). However, in the coding of the live situation and the videotaped material (T_1-T_2), statistically significant differences appeared in all clusters. The same variation between teachers that was noted in inter-coder agreement appeared also in this reliability component.

Speech audibility may have varied for the two teachers between the live situation and the videotaped material. Also, the consistency of the observed features of behavior was reflected in the coding differences.

On the other hand, the coders might have learned to listen for and observe the reactions of the live target.

The mean coefficient differences in *between-coder constancy* were highly significant in all clusters, and these differences were greater when the coding circumstances varied (T_1-T_2). The differences in the level of mean coefficient values varied between teachers and by cluster, as in other reliability components, but in this case the variation was even more outstanding.

Consequently, two different teachers (a man and a woman) as the targets of observation seemed to cause systematic differences in reliability coefficients. The levels of within-occasion reliability (agreement) and between-occasion reliability (constancy) differed considerably, and the consistency of the observed behavior was reflected in a systematic way by cluster.

TABLE 22 Analysis by content, *teacher*: inter-coder agreement. Significance of differences in means of Scott's Pi coefficients and ANOVA by cluster (I, II, III) and by occasion (T_1, T_2, T_3) (n=180, $p<.01$)

	Teacher 1		Teacher 2		Total		Difference df=358	ANOVA df=1/358
	X	SD	X	SD	X	SD	t	F
Occasion T_1								
Cluster I	.53	.17	.61	.16	.57	.17	4.58	21.00
Cluster II	.56	.26	.66	.24	.61	.26	3.59	12.89
Cluster III	.79	.19	.71	.34	.75	.28	-2.73	7.47
Occasion T_2								
Cluster I	.55	.18	.66	.16	.61	.18	5.99	35.89
Cluster II	.67	.25	.75	.16	.71	.22	3.44	11.87
Cluster III	.85	.14	.70	.48	.77	.36	-4.06	16.50
Occasion T_3								
Cluster I	.57	.21	.65	.16	.61	.19	3.93	15.47
Cluster II	.51	.42	.67	.28	.59	.36	4.35	18.91
Cluster III	.62	.60	.58	.58	.60	.59	-.68	.45

TABLE 23 Analysis by content, *teacher*: coder constancy. Significance of differences in means of Scott's Pi coefficients and ANOVA by cluster (I, II, III) and by occasion (T_1-T_2, T_2-T_3) ($p<.01$)

	Teacher 1		Teacher 2		Total		Difference	ANOVA
	X	SD	X	SD	X	SD	t	F
Within-coder Constancy T_1-T_2								
Cluster I	.63	.14	.70	.14	.66	.15	3.18	10.12
Cluster II	.50	.31	.68	.22	.59	.28	4.03	16.25
Cluster III	.80	.18	.44	.60	.62	.48	-4.91	24.12
n=72							df=142	df=1/142
T_2-T_3								
Cluster I	.69	.14	.74	.12	.71	.13	2.39	5.72
Cluster II	.62	.35	.71	.26	.66	.31	1.80	3.24
Cluster III	.72	.47	.66	.48	.69	.47	-.78	.61
n=72							df=142	df=1/142
Between-coder Constancy T_1-T_2								
Cluster I	.42	.20	.60	.15	.54	.19	8.83	77.98
Cluster II	.46	.33	.65	.24	.56	.30	8.77	76.92
Cluster III	.79	.17	.43	.59	.61	.47	-11.22	125.80
n=360							df=718	df=1/718
T_2-T_3								
Cluster I	.55	.19	.64	.17	.59	.19	6.59	43.48
Cluster II	.57	.36	.67	.26	.62	.32	4.43	19.62
Cluster III	.70	.47	.55	.55	.62	.52	-3.91	15.32
n=360							df=718	df=1/718

Grade level: An overview of the mean Scott's *pi* coefficient values and standard deviations in the three clusters indicates systematic variation by grade level. *Inter-coder agreement* mean values (Table 24) were noticeably higher in the coding of the upper level than in that of the middle and lower levels. In Cluster III, differences of mean values were noted between the lower and middle levels. The coefficients were again lowest in the live situation (T_1), and highest in the first coding occasion of the videotaped material (T_2).

Statistically significant differences of means of inter-coder agreement values were found between the three grade levels in Cluster I, between the upper level and other levels in Cluster II, and between the lower and the middle levels in Cluster III.

The differences in *within-coder constancy* (T_2-T_3) between the mean coefficient values of the lower, middle and upper levels were not found to be statistically significant in any cluster (Table 25). However, in the live situation and the first videotaped coding occasion (T_1-T_2), statistically significant differences were found between the lower and middle level mean coefficient values in Cluster I and again between the middle and upper levels in Cluster II.

Statistically significant differences were found in *between-coder constancy* in the live situation and in both videotaped coding occasions (T_1-T_2 and T_2-T_3). These differences existed between the lower and upper levels as well as between the middle and upper levels in Cluster I and in Cluster II, and between the lower and middle as well as the middle and upper levels in Cluster III.

TABLE 24 Analysis by content, *grade level*: inter-coder agreement. Significance of differences in means of Scott's Pi coefficients and ANOVA by cluster (I, II, III) and by occasion (T_1, T_2, T_3) (n=120, p<.01)

Cluster	Lower		Middle		Upper		Total		Differences df=238			ANOVA df=2/357
									1-2	1-3	2-3	
	X	SD	X	SD	X	SD	X	SD	t	t	t	F
T_1												
I	.56	.16	.56	.18	.60	.17	.57	.17	.03	2.07	1.90	2.58
II	.58	.23	.58	.30	.67	.22	.61	.26	-.01	3.04	2.59	4.84
III	.69	.27	.78	.31	.79	.25	.75	.28	2.49	3.02	.25	4.96
T_2												
I	.58	.15	.59	.21	.65	.18	.61	.18	.06	3.33	2.76	6.01
II	.68	.18	.66	.29	.79	.13	.71	.22	-.67	5.46	4.51	13.44
III	.70	.44	.80	.28	.82	.33	.77	.36	2.25	2.46	.43	16.50
T_3												
I	.59	.21	.59	.19	.66	.15	.61	.19	-.12	2.82	3.11	5.28
II	.55	.37	.55	.40	.67	.20	.59	.36	-.15	2.78	2.80	4.77
III	.43	.77	.74	.36	.63	.52	.60	.59	4.02	2.53	-1.91	8.98

TABLE 25 Analysis by content, *grade level*: coder constancy. Significance of differences in means of Scott's Pi coefficient and ANOVA by cluster (I, II, III) and by occasion (T_1-T_2, T_2-T_3) (p<.01)

Cluster	Lower		Middle		Upper		Total		Differences			ANOVA
									1-2	1-3	2-3	
	X	SD	X	SD	X	SD	X	SD	t	t	t	F
Within-coder Constancy												
T_1-T_2												
I	.69	.11	.62	.16	.68	.16	.66	.15	-2.69	-.55	1.78	3.48
II	.60	.20	.51	.34	.67	.21	.59	.28	-1.37	1.56	2.76	4.05
III	.51	.60	.75	.32	.60	.46	.62	.48	2.40	.81	-1.82	3.03
n=48										df=94		df=2/94
T_2-T_3												
I	.71	.13	.72	.13	.71	.14	.71	.13	.47	.18	-.29	.11
II	.61	.34	.66	.30	.72	.28	.66	.31	.69	1.76	1.13	1.61
III	.58	.61	.83	.27	.66	.45	.69	.47	2.54	.69	-2.25	3.49
n=48										df=94		df=2/94
Between-coder Constancy												
T_1-T_2												
I	.55	.16	.51	.22	.57	.16	.54	.19	-2.43	1.41	3.62	7.53
II	.55	.26	.48	.36	.65	.25	.56	.30	-2.46	4.47	6.17	21.39
III	.49	.56	.74	.31	.59	.47	.61	.47	6.15	2.01	-4.32	18.61
n=240										df=478		df=2/717
T_2-T_3												
I	.57	.18	.57	.20	.63	.18	.59	.19	-.01	3.46	3.30	7.45
II	.58	.32	.60	.34	.69	.28	.62	.32	.56	3.92	3.18	8.09
III	.52	.65	.75	.33	.59	.49	.62	.52	4.91	1.33	-4.18	13.01
n=240										df=478		df=2/717

Subject area: An overview of the mean Scott's *pi* coefficient values and standard deviations in the three clusters would appear to indicate systematic variation by subject area.

In Cluster I, the level of *inter-coder agreement* (Table 26) was lower in gymnastics and apparatus than in rhythmic movement-expression and ball games, while in Cluster III the case was exactly the opposite. In Cluster II, the mean coefficient values were higher in apparatus and ball games than in gymnastics and rhythmic movement-expression. Statistically significant differences were found between these subject area mean values in all clusters (I,II,III) and in all coding occasions (T_1, T_2 and T_3), most frequently when ball games and gymnastics were compared with the other subject areas. These differences may be due in part to the constancy variations of the subject area. The differences were reflected in a systematic way, varying according to clusters. Variation, however, was least in Cluster I.

Within-coder constancy was not found to be so sensitive to subject area variation as inter-coder agreement. Statistically significant differences between the mean coefficient values were found in Cluster II between apparatus and gymnastics, and between apparatus and rhythmic movement-expression (Table 27).

This was also true of the repeated coding occasions (T_2-T_3), which indicates the difficulty the coders had in interpreting and coding in a consistent manner, movement and social access variation in gymnastics and rhythmic movement-expression. Apparently, variations in activity/passivity and in the degree of pupils' freedom in social activity were smaller and more clearly defined in ball games and apparatus than in gymnastics and rhythmic movement-expression.

The level of the coefficients in apparatus was higher than in other subject areas. The same difference could be noted in the coding of the live situation and the videotaped material (T_1-T_2). In addition, in Cluster III the ball game mean values were found to be much lower than the mean values of other subject areas. These differences appeared to be statistically significant. Some features of the game situation, such as social form, were obscured in the recorded material.

Statistically highly significant differences were found in *between-coder constancy* both in coding occasions (T_1-T_2) and (T_2-T_3) in all clusters. The variation had the same characteristics as the variation in inter-coder agreement, but was even more pronounced.

The coding content constancy varied according to teacher, grade level and subject area, differing by cluster. In all the coding contents, the between-occasions reliability (constancy) was higher than the within-occasion reliability (agreement). Thus, it was shown that the lack of reliability does not mean that the majority of classifications occurred by chance. The coders' individual and unique manner of interpreting the situation and using the metalanguage of the coding system might have been the main factors causing disagreement.

TABLE 26 Analysis by content, subject area: inter-coder agreement. Significance in means of Scott's Pi coefficients and ANOVA by cluster (I, II, III) and by coding occasion (n=90, <.01)

Cluster	Gymnastics		Apparatus		Rhyt.m. express		Ball Games		Total df=178		Differences df=3/356						Anova
											1-2	1-3	1-4	2-3	2-4	3-4	df=3/356
	X	SD	X	SD	X	SD	X	SD	X	SD	t	t	t	t	t	t	F
T_1																	
I	.54	.17	.53	.18	.60	.18	.63	.13	.57	.17	-.33	2.28	3.84	2.56	4.10	1.26	7.13
II	.58	.26	.63	.30	.54	.25	.69	.20	.61	.26	1.23	-1.07	3.41	-2.24	1.73	4.70	6.46
III	.86	.19	.75	.31	.76	.14	.64	.38	.75	.28	-3.04	-4.22	-4.88	.35	-1.98	-2.70	9.76
T_2																	
I	.57	.19	.56	.15	.62	.19	.68	.17	.61	.18	-.41	1.67	4.15	2.24	5.04	2.36	8.91
II	.64	.24	.78	.17	.67	.23	.73	.19	.71	.22	4.63	.96	3.65	-3.63	-.93	2.67	9.43
III	.86	.16	.80	.31	.82	.15	.62	.58	.77	.36	-1.53	-1.77	3.68	.45	-2.53	-3.04	7.93
T_3																	
I	.58	.18	.58	.16	.62	.21	.66	.20	.61	.19	.09	1.46	2.64	1.47	2.73	1.07	3.30
II	.54	.40	.77	.15	.45	.39	.61	.37	.59	.36	5.24	-1.50	1.23	-7.40	-3.92	2.81	14.44
III	.86	.19	.65	.48	.56	.50	.34	.86	.60	.59	-3.91	-5.39	-5.60	-1.24	-2.98	-2.08	13.31

TABLE 27 Analysis by content, subject area: coder constancy. Significance of differences in means of Scott's Pi coefficients and ANOVA by cluster (I, II, III) and by coding occasion (T1-T2, T2-T3) ($p < .01$)

Cluster	Gymnastics		Apparatus		Rhyt.m. express		Ball Games		Total df=178		Differences df=3/356						Anova
											1-2	1-3	1-4	2-3	2-4	3-4	
	X	SD	X	SD	X	SD	X	SD	X	SD	t	t	t	t	t	t	F
Within-coder Constancy																	
T_1-T_2																	
I	.70	.11	.63	.17	.64	.18	.68	.11	.66	.15	-2.19	-1.72	-1.00	.26	1.40	.90	1.86
II	.53	.29	.68	.27	.52	.25	.65	.30	.59	.28	2.34	-.17	1.72	-2.69	-.52	2.01	3.26
III	.76	.31	.73	.41	.70	.19	.29	.70	.62	.48	-.27	-.98	-3.71	-.48	-3.33	-3.42	9.15
N=36														df=70			df=3/140
T_2-T_3																	
I	.72	.10	.68	.14	.71	.15	.74	.14	.71	.13	1.52	-.27	.76	1.00	1.97	.86	1
II	.57	.27	.79	.18	.56	.36	.72	.34	.66	.31	4.07	-.18	2.00	-3.50	-1.14	1.93	5.30
III	.74	.36	.75	.43	.64	.40	.61	.66	.69	.47	.10	1.11	1.06	-1.12	-1.09	-.25	.88
N=36														df=70			df=3/140
Between-coder Constancy																	
T_1-T_2																	
I	.53	.17	.49	.20	.55	.21	.59	.16	.54	.19	-.63	1.33	4.06	2.72	5.72	2.23	9.80
II	.49	.32	.64	.30	.47	.27	.63	.27	.56	.30	4.67	-.36	4.62	-5.41	-.30	5.43	16.82
III	.72	.31	.73	.36	.69	.18	.29	.70	.61	.47	.25	-1.17	-7.71	-1.31	-7.60	-7.57	44.85
N=180														df=358			df=4/716
T_2-T_3																	
I	.55	.18	.55	.16	.61	.20	.66	.18	.59	.19	-.16	2.80	5.58	3.11	6.12	2.45	14.73
II	.52	.35	.76	.16	.55	.35	.67	.32	.62	.32	8.44	.83	4.31	-7.43	-3.33	3.47	24.30
III	.72	.36	.69	.42	.60	.40	.47	.75	.62	.52	-.59	-2.91	-3.97	-2.12	-3.45	-2.03	8.66
N=180													df=358				df=4/716

6.3.2 Reliabilities of individual categories

The inter-coder agreement was assessed for various individual categories of the three clusters of the PEIAC/LH-75 by using the Kendall coefficient of concordance, *W* (Siegel 1956). In the statistical processing of the material, the sub-program FORTRAN NMCC was applied. The total percentage of frequencies, summed per category and per observer over a sample of 24 lessons, was ranked separately by the categories of the three clusters and by occasions T_1, T_2 and T_3. A Chi Square test was used for estimating the degree of the statistical significance of the coefficients (Table 28).

The intraclass correlation coefficient was also estimated for each category of observation from the variance between a sample of 24 lesson observations and the variance between the six observers' percentage per category, separately by cluster and by occasions. The stability estimates were not computed in connection with these indices, but the range of variation of indices between coding occasions gave a rough description of the inter-coder agreement stability by individual categories and by the clusters of PEIAC/LH-75.

As can be seen in Table 28, the inter-coder agreement was rather high, 23 of the categories yielded a value of *W* statistically significant at the .01 level. Only the indices of one category with low frequencies (I/03), and the categories indicating a confused situation, also occurring infrequently (I/12, II/8 and III/7), were not statistically significant in all occasions. There was also one category, which all observers did not use in the first coding occasion (III/6), and a *W* could not be computed in this case. The significant value of *W* means that the six independent observers were applying essentially the same standards in ranking the sample of 24 lessons by using most of the categories of the system. However, as cited earlier, a significant value of *W* does not mean that the rankings observed are correct. In this special case, because a relevant external criterion does not exist, the ranking of lessons by categories served more or less as an "objective one" (cf. Siegel 1956).

The level of coefficient values varied between clusters in accordance with the level of overall reliability determined earlier by computing Scott's *pi* (see Table 18). Analyzing the values of videotaped material observation in occasions T_2 and T_3, it was noted that the general level of reliability of the individual categories was highest in the Social Form Cluster III, Md .95 and Md .89, second highest in the Movement and Social Access Cluster II, Md .86 and Md .73 and lowest in the Verbal Cluster I, Md .72 and Md .72. Inter-coder agreement also diminished with time, and most strongly in Cluster II, whereas in Cluster I it remained at the same level in both occasions.

In comparing the *W* values obtained in different situations, it was noted that inter-coder agreement was higher in the live situation than in the videotaped material observation in Cluster I, Md .76-.72 and in Cluster III, Md .96-.95, whereas in Cluster II it was higher in TV, MD .81-.86.

When the variation of means (see Table 18, p. 135) was tested by Scott's *pi*, the opposite situation was found to be true in Cluster I. It is possible that these differences of *pi* and *W* values reflect the role of chance agreement. (see Table 28, p. 125) As cited earlier, Scott's *pi* describes the average of observer

agreement about the proportions of behaviors in the categories, corrected for chance agreement.

It can also be seen in Table 28 that the level of the intraclass correlation coefficient was in general rather high, but lower than the values of the coefficient of concordance, *W*. The variation of the level of this coefficient was also generally in accordance with the variation of *W*, and, in categories occurring frequently, the difference between indices was very small. Intraclass correlation possesses a known sampling distribution and, therefore, it may be given a standard psychometric interpretation. In this case, when the correlation coefficient was computed from mean squares obtained from the six observers' percentages per category by cluster, high values indicate that the variation between observers was small relative to the variation among observations in the sample of 24 lessons used in the study. The intraclass correlations were sensitive to variations of marginal frequencies, which were also noted in analyzing the variance of the means of Scott's *pi* coefficients for determining the level of the objectivity of coding.

Inter-coder agreement on the frequencies was satisfactory, although category I/03, with low frequencies, and the confused situation categories diminished the level of overall reliability decisively. Thus, it can be concluded that the three dimensional measuring instrument PEIAC/LH-75 was reliable when estimated by using a nonparametric coefficient of concordance, *W*. However, some revisions are needed. The question of inter-coder agreement is further examined in the following section using discriminant analysis techniques.

6.3.3 Discussion of overall reliability results

In this section the general problems of reliability related to the procedures of categorization are discussed. The coefficients obtained in these analyses can be compared with reliabilities obtained in other studies. According to Flanders (1967b, 166) a Scott's coefficient of .85 or higher is a reasonable level of performance. This statement is based on the analysis of errors of two observers during a four-month period, in which the original 10-category system was used. As a rule, however, in studies using instruments modified and expanded from the Flanders system, coefficients have failed to reach the limits suggested by Flanders (Hough & Ober 1967, 334). It was also noted by Flanders (1970) that by using a subdivided FIAC system the reliability checks produced inter-coder coefficients between .70 and .86, and during a "difficult" observation, .56.

TABLE 28 Analysis categories: Kendalls' W, interclass correlation and Chi Square -test computed by categories in clusters I, II and III of the PEIAC/LH-75 and by coding occasion (T_1, T_2, T_3)

Cluster		Categories	T_1 (live situation) N=24 %	W	Intrac. correl.	x_2 df 23	T_2 (video re. obs. 1.) N=24 %	W	Intrac. correl.	x_2 df 23	T_3 (video rec. obs. 2) N=24 %	W	Intrac. correl.	x_2 df 23
		Teacher's talk, movement, pupils' talk, other												
I	Teacher	01. Accepts, praises	4.5	.81	.78	112.2	3.1	.79	.75	109.4	3.1	.73	.68	100.9
		02. Gives corr. feedback	5.1	.60	.52	82.8	5.6	.64	.56	87.6	4.4	.63	.56	86.9
		03. Uses ideas dev. by pup.	0.8	.33	* .19	45.1	0.3	.36	.24	50.3	0.4	.38	.25	52.0
		04. Asks, init., term. act.	8.6	.81	.77	112.0	6.7	.88	.86	122.0	7.2	.90	.88	123.9
		05. Presents inform., org.	37.6	.80	.76	110.1	39.6	.83	.79	113.9	42.1	.77	.72	105.9
		06. Gives dir., comm.	4.3	.55	.46	75.5	3.8	.71	.65	98.1	3.1	.71	.65	98.1
		07. Criticizes	1.3	.74	.69	102.4	0.8	.69	.63	95.1	0.8	.70	.64	96.4
	Pupil	08. Answers questions	0.8	.57	.48	78.7	0.6	.62	.54	85.7	0.8	.67	.60	92.1
		09. Speaks spontan., init.	1.9	.79	.74	108.4	1.7	.72	.67	99.6	1.3	.79	.74	108.5
	Teacher	10. Silent guidance	28.1	.86	.84	119.2	30.9	.88	.85	121.5	30.0	.88	.86	121.5
		11. Silent participation	6.1	.94	.93	130.1	5.8	.96	.95	132.0	5.8	.96	.95	132.5
	Other	12. Confused situation	1.4	.35	.22	48.5	1.1	.08	* -.11	10.9	1.0	.26	* .11	35.2
II		Pupil's collective movement activity/ passivity and social access												
	Activity	1. Contacts, ideas cont.	14.8	.88	.85	120.7	11.4	.90	.88	124.4	10.5	.73	.67	100.0
		2. Contacts free, ideas cont.	37.7	.92	.90	126.2	40.7	.95	.95	131.6	42.3	.87	.84	119.3
		3. Contacts free, ideas open	9.9	.96	.95	131.8	8.1	.93	.91	128.1	8.0	.85	.82	117.8
		4. Pupils' spont. activity	0.6	.48	.38	66.8	0.4	.51	.41	70.0	0.3	.37	.25	51.7
	Passivity	5. Pupils follow instruction	25.9	.89	.86	122.2	27.2	.91	.89	125.0	27.9	.89	.87	123.3
		6. Pupils organization	8.9	.73	.67	100.6	10.5	.81	.77	111.4	9.5	.72	.66	98.8
		7. Pupils wait for turn	1.0	.41	.29	56.5	0.6	.50	.40	69.0	0.4	.36	.23	49.8
	Other	8. Confused situation	1.2	.46	.35	63.2	1.1	.16	* -.01	21.5	1.1	.37	.24	50.5
III		Social form												
	Situation	1. Complete class, uniform task	31.3	.97	.96	133.5	31.5	.96	.95	132.2	11.9	.96	.95	131.8
		2. Divided class, uniform task	27.3	.97	.96	133.2	28.0	.96	.95	132.2	29.1	.89	.87	122.4
		3. Divided class, different tasks	23.4	.98	.98	135.8	22.6	.95	.94	131.6	22.4	.92	.90	126.0
		4. Div. cl. diff. task within gr.	8.9	.96	.95	132.7	9.0	.97	.97	134.5	8.4	.92	.91	127.0
		5. Indivudual work, unif. tasks	7.3	.93	.92	128.4	7.4	.93	.92	128.3	6.8	.78	.74	108.2
		6. Individual work, diff. tasks	0.3	-	-	138.0	0.3	.71	.65	98.0	0.2	.67	.60	92.0
		7. Other, conf. situation	1.5	.58	.50	80.2	1.3	.63	.56	87.8	1.2	.19	* .03	26.1

all coefficients concordant to the level of significance 1 %
* = p > 0.05 significance and beyond
6 observers, 24 lessons, 4800 time units., tot. 28800 time units

Using a unidimensional 16 category adaptation from Flanders' FIAC in physical education, Splinter (1980, 111) fixed the criterion of reliability at Scotts' Pi .70. Also when using multidimensional observation instruments modified from the Flanders system and constructed for the observation of physical education classes, the level of performance was lower. Gasson (1971), in analyzing the verbal and nonverbal behavior of the teacher and pupils and the location of the teacher in relation to the class, noted that an inter-coder coefficient of at least .70 for each of three dimensions would be Bookhout (1967), in his multidimensional observation instrument, accepted the level of .40 reliability in selecting variables to be submitted to factor analysis on the basis and stated that the higher the reliability cut-off point set, the fewer variables would be submitted and the greater the risk of throwing away valuable data.

However, Barrett (1971) recommended a level of 90% for determining the objectivity of coding by computing the percentage of inter-coder agreement among trained observers for a multidimensional system developed primarily as a research tool.

In the present study, in which a three-dimensional category system was used, the level of inter-coder agreement Scott's Pi was rather low, Md .65, varying between clusters as follows: Cluster I, .61, Cluster II, .65, and Cluster III, .69, e.g. in the observation of video-taped material (T_2). The reliability index used here, Scott's *pi*, took into consideration the estimated role of chance, and was roughly interpreted in this context to indicate the extent to which the codings of the six observers exceeded chance agreement divided by the amount that perfect agreement would exceed chance. However, chance seemed to have less significance as an error-causing factor than the coders, coding target and coding occasions. The general character of errors was found to be more systematic than random.

As was expected, within-occasion reliability Scott's Pi (agreement) (Md .61) was lower than between-occasion reliability (constancy) (Md .64). In Cluster I, this difference (.61-.71) was found to be systematic. In a comparison *pi* values, a wide variance was evident in inter-coder agreement by coder pair (T_1, .45-.65, T_2, .46-.73, and T_3, .43-.72) and still wider in coder consistency by coder pair (T_1-T_2, .37-.68, T_2-T_3, .41-.72). A similar range of variation was not evident in within-coder constancy which ranged between .64-.68 (T_1-T_2) and .69-.78 (T_2-T_3).

The coders' interpretations of the situations and use of the metalanguage of the coding system were unique. Regarding the coding occasions, inter-coder agreement diminished with time (T_2-T_3), except in the verbal cluster where the level of inter-coder agreement remained at the same level. It was apparent that the differences between coders were somehow compensated for in this cluster. The group of observers was heterogeneous with some demonstrating a higher level of agreement with themselves, whereas others agreed more consistently with other observers than with themselves. This kind of change phenomenon was also found by Barrett (1971), and has relevance to observer training as well as to a continuous estimation of reliability and objectivity. The checks of observer agreement carried out at the end of the training period or at given intervals were not enough to avoid systematic errors in coding. However as Komulainen pointed out in analyzing the overall reliability of an observation

instrument modified from the Flanders category system, "constancy control through time must also be resorted to" (Komulainen 1970, 12).

Cluster I was modified and expanded from the Flanders category system and therefore the coding system proposed here uses the same principles of categorization and training procedures. Two of the ground rules given to trainees to increase consistency when choices occurred, need to be discussed here in more detail. First, the rule, "always maximize information by choosing the least frequently occurring category, when there is a choice", and second, "if the observer feels that the pattern at the moment is restrictive, he is cautious in the use of direct categories, but he remains alert to a shift in momentary patterns by remaining alert to the total social situation" (Flanders 1967b, 159).

The results obtained in this study with a modified instrument and six-second time intervals seem to confirm that these ground rules are an invitation to biased observation. However as Flanders has stated that there is a theory of the "unbiased, biased observer", "which contends that even if the observer's assessment appears to be biased, he is unbiased in that he remains open to all evidence of a changing situation" (Flanders 1967b, 159). It is evident, too, that the time interval of six-seconds caused additional problems of choice in Cluster I. This error-causing effect was found to be present in the results, judging both by the level of the coefficients in different clusters and by the number of categories in these cluster comparisons. It is advisable to take this into account, however, as Flanders (1970) points out, by choosing time intervals as the unit of analysis: "When such time intervals are small, compared with the cycles or natural units which are of interest, then not too much error is introduced. This approach has the advantage that the observer does not have to make snap judgements about the beginning and end of natural units while he is observing" (Flanders 1970, 164).

By using three-second time interval frequencies, as in the FIAC, compared with the six-second interval used in the PEIAC/LH-75, the frequencies are naturally higher and are also reflected in the level of reliability. In the other clusters, the range of variation of frequencies and also the *pi* coefficients were higher, and the role of random errors greater. The use of categorizing principles merits closer examination in connection with different time intervals.

One factor contributing to the unsystematic variation of reliability components in the Movement and Social Access Cluster II and in the Social Form Cluster III was related to the videotape recording and to the quality of the videotapes used. On several occasions, the video segment was less than adequate, with either teacher or student behavior obscured from view. It may be that the camera angle was not sufficiently thought out with the observation of these activities in mind. In general, the recording was found to have a more detrimental effect on visibility than on audibility. The rules guiding videotape recording and categorizing principles also merit a closer look.

The coding errors caused by the constancy and nature of the coding targets (teacher, grade level, subject area) was rather more systematic than random, and was reflected differently in different clusters. The reactions of "living instrument" to "living target", such as teachers, were clearly visible. When comparing grade level effects and teacher effects on the level of

reliability, Tavecchio (1977, 95) noted that the results obtained in a study using the generalizability of scores and profiles for reliability assessment, seemed to confirm the view that the former is "nested within teacher". This was found to be a general characteristic also of the present study because within-coder constancy variation was not statistically significant by grade level in any clusters as it was by teacher. As the coders became acquainted with the coding target, random errors became a systematic way of interpreting teaching behavior individually and uniquely, according to the coding system.

It was also evident that there were common elements and a certain degree of consistency in the interaction pattern in the condition of different P.E. subject areas. This consistency of variation seems to be reflected in the results of this examination, as well as in Cheffers' (1973) study where reliability was determined by submitting cell rankings to analysis using Kendall's coefficient of concordance, *W*.

One qualification is necessary here, however. There were various kinds of errors by individual coders, although no attention was paid to the meaningfulness of errors in this study. An examination of the variance of coders would be a first step toward this kind of study.

Thus, it can be concluded that there was a high degree of consistency both in coding behavior and in the target observation. The results obtained suggest that the theory of the generalizability of scores and profiles presented by Cronbach et al. (1972), in which the question of reliability resolves into a question of the accuracy of generalizability, merits consideration in examining the multidimensional problems of reliability and validity in the construction of measuring instruments for the observation of physical education classes.

The consistency among samples of behavior challenges the investigators to work out concepts of variables to be measured as a part of instrument validation as well as a study of instrument precision (McGaw et al. 1972).

6.3.4 Summary of the reliability and objectivity of coding

The aim of this part of the study was to identify and describe the methodological problems involved in an observation instrument proposed for analyzing the directive/non-directive aspects of interaction in physical education teaching (Heinilä 1977a). This inquiry was designed to investigate and assess the within-occasion reliability (agreement) and between-occasion reliability (constancy) by cluster, coder pair by situation and content of lessons (see page 84 and 112).

The overall reliability was determined by cluster, using the scores of six trained observers, each having observed 24 P.E. lessons (20 minutes each) three times, on occasions randomly placed at one-month intervals, first in a live situation and then twice more in videotaped situations. The reliability of the different clusters was assessed by using the profile method, and was computed by using Scott's *pi* coefficient (Scott 1955). A total of 8424 Scott's coefficients were computed. The coefficients were examined by means of t-tests and a one-way analysis of variance (ANOVA). The variation of reliability coefficients was examined by analyzing the between-coder reliability (agreement) and the

within-coder reliability (constancy). The contribution made to variation by the different components was analysed by means of a one-way analysis of variance. The reliability of individual categories was also determined by using the nonparametric Kendall coefficient of concordance, *W* and also by computing intraclass correlation coefficients.

Summarizing the main results, the average level of mean coefficient values was rather low and varied according to cluster (pi I/.61, pi II/.65, pi III/.69) and reliability component (inter-coder agreement .65, within-coder constancy .69, and between-coder constancy .60), as indicated in results of the videotape recordings T_2 and T_3. The range of variation and dispersion of coefficients was high.

In Cluster I, these "errors" were found to be more or less systematic in character. The reliability index used, Scott's *pi* coefficient, took into consideration the estimated role of chance in determining the level of reliability. However, in connection with the sample used in this study, chance seemed to have less significance as a reliability-decreasing factor than that resulting from the coders, coding occasions and coding target. The chance phenomenon that was found to occur in the use of the categorizing principles of Cluster I judging both the between-coder and within-coder constancy comparisons, seems to have relevance both to the development of the structure of the measuring instrument and to the improvement of the general rules guiding the coders and the training of observers.

The reliability, operationalized as inter-coder agreement and assessed by means of the Kendall coefficient of concordance *W*, was found to be rather high. Twenty-three of the 27 categories yielded a value of *W* significant at the .01 level (Chi Square test) but in all coding occasions, the coefficients of four categories of infrequent occurrence (I/03, I/12, II/8 and III/7) were not statistically significant, as was also evident from the computed coefficient of concordance. Thus, the categorizing principles need to be considered more closely.

In addition to the assessment of the objectivity of coding, the information concerning the "consensual ordering" of lessons by individual categories may be useful for refining the structure of the instrument and the rules of categorization to facilitate the measurement of theoretically important concepts. It can be concluded, after reviewing the results of this examination of both the overall reliability and the reliability of individual categories of the measuring instrument, that more information is needed about general factors causing errors in coding before the category system can be implemented to objectively measure these concepts.

6.4 Variability in coding

This section will concentrate on the methodological problems associated with the development of the observational instrument and report on an experiment

made to examine more closely the sources of variability in coding by means of the multiple discriminant analysis technique (Heinilä 1980).

As Dunkin and Biddle (1974, 78) noted, when reviewing approximately 500 descriptive studies dealing with the observation of classroom interaction, "the terms reliability and validity have technical meanings when used to describe instruments for measurement of teaching" and "to say that the instrument is reliable means that it provides the same score of measurement for repeated applications to the same teaching events", and "to say that an instrument is valid means that it measures what we think it is measuring".

In most cases the investigators constructing observational instruments consider only observer agreement and neglect the study of validity. This has been common to researchers dealing with observation of physical education teaching and applying modified instruments already validitated, such as the most commonly used Flanders FIAC system (e.g. Dougherty 1970, 47, Gasson 1971, 38, Mancuso 1972, 84-85). But as we know, an instrument may be reliable without being valid but not vice versa, and thus it is appropriate to concern ourselves also with the crucial question of variability of coding in connection with physical education studies.

As note earlier, because in observation studies the observer and the classification system together form the measuring instrument, reliability is not to be regarded only as a property of an instrument but as that of measurement. Similarly, an instrument itself is neither reliable nor unreliable, and reliability can be judged only when it has been used to collect data and data have been manipulated in some way to produce scores. Thus, reliability has to do with the reliability of scores the observer becomes an additional source of errors of measurements. According to Komulainen (1973, 12), "the value of the final results depends crucially on the accurate use of the metalanguage of classification system in the coding process". Therefore the main problem in developing and observational instrument is how to get adequate information for refining the classification system and especially the rules guiding the observers so that theoretically important concepts could be measured objectively (see Flanders 1970, Komulainen 1970b, 24).

Because there was no external criterion available to assess the validity of these codings it was decided to use multiple discriminant analysis for examining more closely the variability of coders, i.e. to describe common features of disagreement.

6.4.1 Research task related to the variability of coding

The purpose of the study on the variability of coding was to determine the degree of variability in the codings of different observers when using the categories of the three cluster category system PEIAC/LH-75 (Heinilä 1977a).
In this connection the aim was:

- to find those discriminant functions that best separate the observers from each other, in other words, maximize the between-observer variance, relative to the within-observer variance,

- to describe factors connected with the use of the category system that cause such differences and thus reduce the degree of agreement among coders,
- to examine the structure of the observer group in terms of the noted deviations and thereby attempt to describe the degree of validity in this "testing".

6.4.2 Discriminant analysis of the observational data

Although discriminant analysis has rarely been used in observational studies, it is appropriate to explore its applicability as a method of assessing and describing factors predicting inter-coder dissagreement. This method is presented more comprehensively by Cooley and Lohnes (1971). Here the main features, tasks, assumptions and principles of the interpretation of results will be considered.

The objective in discriminant analysis is to find a linear combination of the independent variables that minimizes the probability of misclassification of the individuals or objects into their respective groups. In the two-group discrimination problem the attempt is to find a single linear composite of the predictor variables what could discriminant between the groups. It can be obtained by maximizing the ratio of between-groups to within-groups variability and under certain conditions, it product the smallest missclassification error rates. Thus, the single linear composite yields a new axis along which the groups are maximally separated. In multiple discriminant analysis, when more than two groups are involved; the situation is much the same. For optimality, we must assume (1) multivariate normality of the predictor variables and (2) equal variance-covariance matrices in each of the groups. (Cooley and Lohnes 1971, 245, Dillon and Goldstein 1984, 395).

In this analysis the codings (scores) of different observers may be regarded as forming criterion groups (A-F), representing the universe of observers using the category system (27 categories) in the universe of coding situations (n=144) (see Appendix 3). In discriminant analysis a linear function (I) is derived from the predictors (categories of the classification system) so that this function maximally separates the criterion groups (codings of observers). The residuals are treated in the same way. This may result in a new function (II), orthogonal to the former, which improves discrimination of the observer groups. If these functions should prove statistically significant, a curvilinear dependence exists between predictors and criterion.

This model of analysis also makes it possible to classify observers by using discriminant functions in certain groups according to their scores. If we know, for example, the codings of observers A and B and we wish to place them into certain groups according to the set-up below:

Observer is classified into group A or B

		A	B
Observer is a member		right	wrong
of either group		wrong	right

The discriminant analysis makes it possible to minimize the proportion of placement of observers into "wrong" cells.

The assumptions concerning the level of measurement also need to be considered. The discriminant analysis, like the analysis of variance and factor analysis, presupposes that the measurement fulfills the requirements of interval scales. Nevertheless, these methods of analysis have often been used with ordinal data. For example, Cooley and Lohnes (1971) presented a research example by using such data. Such methods have also been used in observational studies by such researchers as Bookhout (1967), Hanke (1980b), Heinilä (1970, 1971, 1980), Komulainen (1973), Medley and Mitzel (1958) and Soar (1968). However, because of the nature of the measurement scale, the interpretation of the results remains largely tentative.

The starting point of the discriminant analysis in the present study was the marginal distributions of categories of the 24 lesson data (T_2) (Appendix 3) as coded by six trained observers, as well as the 27 categories of the three clusters of the classification system. The observation groups were structurally homogeneous and there were differences in the mean distributions of variables. The data fulfilled the requirements set on the number of criterion groups and variables. The use of discriminant analysis was not equally well justified with regard to the level of measurement. This will be taken into account in the interpretation of the data.

6.4.3 Interpretations of the discriminant analysis

The results of this analysis are presented in Table 29, in which are listed five discriminant functions, the maximum number possible since there were originally six groups. The table displays the structure and significance of discrimination. The Chi Square test, computed from Wilks' *lambda*, indicated that of the five discriminant functions separating observers, the first two discriminations were statistically highly significant and the third almost significant. It was further established that the power of the discriminant functions to separate observers was great, since their canonical correlations were relatively high. The first discriminant function proved clearly more powerful than the other four with 56%, the second having only 21%, and the third, 11%. From the point of view of interpretation, the first three discriminant functions were the clearest and theoretically important.

The program selected 13 of the 27 classification categories and set them in sequence according to how much they increased the model's discriminating power. It is possible even on this basis to get an idea of the nature of the

discrimination. The discrimination model included the nine categories of the Verbal Cluster I and four of the Movement and Social Access Cluster II. Both categories, which occurred rarely and those occurring most frequently were represented. In previous studies (Heinilä 1977a), the former categories were found to possess low and the latter high reliability.

6.4.4 Content and interpretation of discriminant functions

The following principles and sequence were used in the interpretation of the contents of the discrimination dimensions: First, note was made of the variables that had obtained high weights on scaled eigenvectors (s) and of their relative discriminating power. Second, it was ascertained how highly discriminant functions correlated (r) with variables selected into the model. Third, it was established how known groups (observers) were placed on the discriminant dimension on the basis of their means and standard deviations on these dimensions. Finally, their mutual placement in the discrimination plane, formed by two discrimination dimensions at a time, was studied.

From the structure of the coefficients, and the nature of the factors, the five functions extracted appear to measure the following variations in the coding behavior of observers A-F:

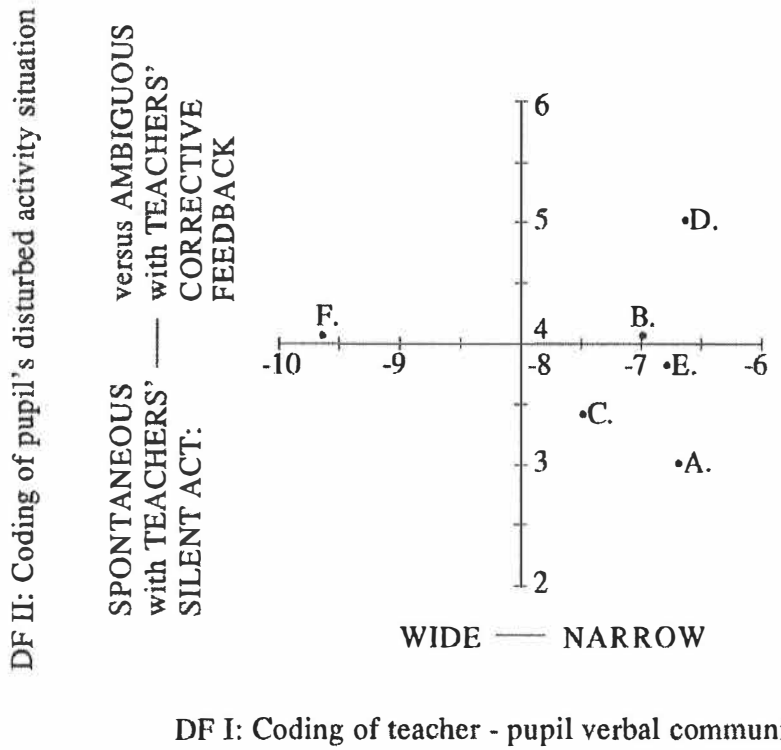
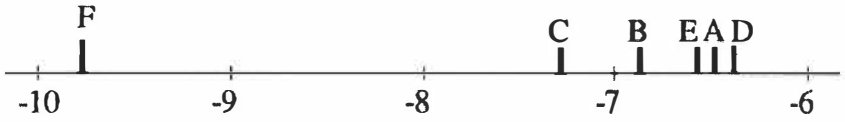


FIGURE 14 Placement of observer A-F group centroids on the discrimination plane formed by discriminative functions I and II

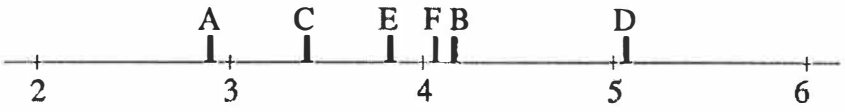
TABLE 29 Discriminant analysis on observers and process variables (PEIAC/LH-75)

Cluster	Categories		Observers A	B	C	D	E	F	Power of discrimination in categories JN	F	n_1	F_{ls}	I Discriminative function s	r	II Discriminative function s	r	III Discriminative function s	r	IV Discriminative function s	r	V Discriminative function s	r
I	01. Accepts, praises		5.46	4.92	5.17	5.62	5.42	10.21	6	5.07	30	4.84	-0.77	-0.51	-0.28	0.10	-0.57	-0.02	0.30	0.24	-0.43	0.31
	02. Gives corr. feedback		6.33	12.54	10.67	16.58	11.03	10.04	5	5.27	25	6.99	-0.18	0.09	0.88	0.56	-0.52	0.07	-0.58	0.23	-0.82	-0.19
	03. Uses ideas dev. by pup.		0.17	0.88	0.21	0.25	1.00	1.25	8	4.59	40	2.61	-0.17	-0.31	0.24	-0.12	-0.75	-0.64	-0.31	-0.18	0.87	0.13
	04. Asks, init., term. act.		11.00	12.67	10.79	11.83	13.42	21.04	3	6.40	15	16.32	-0.99	-0.42	0.33	-0.12	-0.73	-0.23	-0.82	0.10	-0.22	0.21
	05. Presents inform., org.		79.13	72.67	83.92	79.46	88.13	72.25	13	3.48	65	0.77	-0.44	0.20	-0.13	-0.10	-1.13	0.11	-1.19	0.47	-0.24	0.54
	08. Answers questions		1.17	1.08	0.54	0.88	0.92	2.46	10	4.01	50	1.62	0.29	-0.37	-0.94	0.04	-0.36	-0.20	0.54	0.38	-0.25	0.14
	09. Speaks spontan., init.		1.92	1.67	3.58	3.04	1.96	7.79	1	9.98	5	120.18	-1.11	-0.68	0.37	0.15	0.13	0.28	-0.36	0.05	-0.20	0.21
	10. Silent guidance		71.75	73.20	63.00	58.00	57.13	48.03	7	4.83	35	3.51	-1.10	0.29	0.18	-0.24	-1.77	-0.09	-0.59	0.22	-1.47	-0.58
	11. Silent participation		10.71	13.08	11.58	12.25	11.38	10.75	12	3.61	60	1.01	-0.41	0.03	0.55	0.05	-0.98	-0.03	-0.66	-0.03	-0.30	-0.11
II	2. Contacts free, ideas open		85.00	80.95	87.08	81.54	79.83	73.58	9	4.28	45	2.04	0.28	0.09	-0.14	-0.07	0.65	0.11	-0.54	-0.04	0.38	-0.07
	4. Pupils' spont. activity		1.17	0.71	1.33	0.46	0.71	1.21	11	3.77	55	1.30	0.24	-0.06	-0.48	-0.11	0.44	0.10	0.29	-0.00	-0.96	-0.01
	7. Pupils wait for turn		0.92	1.25	2.29	0.33	0.67	2.25	4	5.76	20	10.18	-0.51	-0.37	-0.39	-0.27	0.38	0.21	-0.45	-0.32	-0.48	-0.42
	8. Confused situation		1.96	2.04	2.00	2.63	2.00	2.17	2	7.25	10	32.04	0.15	0.03	0.59	0.63	0.27	0.33	0.16	0.24	0.80	0.11
	Number of observations (144)		(24)	(24)	(24)	(24)	(24)	(24)	$n_2 \geq 138$				56.4 % xxx		21.0 %		10.7 %		7.0 %		5.0 %	
DF I: Coding of verbal communication wide - versus narrow		m	-6.69	-7.01	-7.41	-6.62	-6.80	-9.59	F ≠ others				$x^2_{17} = 100.9$ Rc = 0.73								Of total discrimination	
		d	0.64	0.80	1.06	0.88	0.65	1.62														
DF II: Coding of pupils' activity as spontaneous with teachers' silent activity - versus ambiguous with teachers' corrective feedback		m	3.04	4.05	3.40	5.06	3.87	4.05	D ≠ others						xxx $x^2_{15} = 46.72$ Rc = 0.54							
		d	0.81	0.83	0.84	1.62	0.53	1.02														
DF III: Coding of the sequence of verbal and nonverbal communication by using infrequently occurring - versus frequently occurring categories		m	-9.69	-10.31	-9.16	-9.38	-10.33	-9.86	B and E ≠ C and A, D and F								x $x^2_{13} = 25.74$ Rc = 0.42					
		d	0.67	0.94	0.97	0.89	0.96	1.42														
DF IV: Matter-of-fact-centered coding of teacher talk - versus other		m	-8.07	-8.61	-9.15	-8.47	-9.04	-8.50	E and C ≠ others										$x^2_{11} = 17.42$ Rc = 0.35			
		d	0.93	0.83	0.97	1.10	0.86	1.25														
DF V: Silence-centered coding - versus other		m	-5.59	-6.26	-5.90	-5.63	-5.27	-5.62	B ≠ others												$x^2_9 = 12.64$ Rc = 0.30	
		d	0.98	0.97	1.06	0.95	1.08	0.96														

s = variables scaled in w-metrics, r = correlations, m = means on discriminative function, d = deviations on discriminative function



DF I: Coding of teacher-pupil verbal communication as wide-versus narrow



DF II: Coding of pupil's disturbed activity situations as ambiguous with teacher's corrective feedback- versus spontaneous with teacher's silent activity

DF III: Coding of the sequence of teacher-pupil verbal and nonverbal communication by using infrequently occurring - versus frequently occurring categories

FIGURE 15 Placement of observers A-F controids on the discriminant dimensions I, II and III on the basis of their means and standard deviation on the function

DF I: *Coding of Teacher-Pupil Verbal Communication: Wide versus Narrow.* The first and most important discriminant function distinguished the observers who had made a wide use of the categories of verbal communication from those who had used only some categories. The following categories, besides being highly related to discriminant functions, obtained high weights on scaled eigenvectors: pupil speaks spontaneously (I/09), teacher asks, initiates and terminates activity (I/04), teacher accepts, praises, encourages (I/01). On the basis of the placement of observers on the discrimination dimensions (Figure 15), observer F deviated clearly from the rest, most clearly from observers D and A, and was placed at a distance of over two standard deviations from the others. The observer in question was found to deviate significantly from the others also in the analysis of inter-coder agreement (Heinilä 1977a). The nature of this factor was then examined closer, as was the shift phenomenon by coder, which reduced the index of inter-coder agreement of the whole group. The way in which observer F used the classification system showed a tendency to code more frequently than the others the occurrence of "verbal communication,

teacher and pupil initiative and response". The observer in question also attempted to take into account infrequent and more rapidly occurring events in order to describe the continuity of communication, whereas other observers were content with a less detailed coding of communication.

It is possible that the time interval of six seconds was reflected in these coding differences as well as Rule 4 (see Chapter 5, page 82).

DF II: Coding of Pupils' Collective Activity Situations: such as Ambiguous with Teachers Corrective Feedback versus Spontaneous with Teacher's Silent Activity. This discriminant function separated observers on the basis of how they coded ambiguous situations. An examination of the weights of scaled eigenvectors and of correlation coefficients indicates that the most important categories in this discrimination were the category describing the ambiguity of pupil activity (II/8), teacher's corrective feedback (I/02) and teacher's silent participation in movement activity (I/II). When the placement of observers on the discrimination dimension was analyzed (Figure 14) it was seen that observer D differed clearly from the others, especially from observers A and C. Where observer D tended to code an ambiguous situation using the category "confused situation" (II/8), the rest, and particularly observer A, were more inclined to code it as "spontaneous pupil activity". Similarly, observer D coded the teacher's verbal behavior as "corrective feedback and teacher silent participation" more frequently, while the others used the category "teacher follows pupils' activity, guides silently" (I/10).

It appears that it was difficult to draw a line between confused and spontaneous pupil activity situations.

DF III: *Coding of Verbal and Non-verbal Communication: Infrequently Occurring versus Frequently Occurring Categories.* This third discrimination dimension was not as easy to interpret as the first two dimensions. It was, however, found to be statistically significant and quite interesting from the point of view of the theory and content validity of coding. The discrimination between observers was again related to coding differences in combining non-verbal and verbal communication. For interpretation, the most important discriminating categories proved to be the verbal category "teacher uses ideas, movement tasks suggested by pupils" (I/03) and the category indicative of teacher initiative "teacher asks questions, initiates and terminates activity" (I/04). Included in the model was the most frequently occurring pupil collective activity category "inter-pupil contacts and movement free, range of ideas controlled" (II/2), whose correlation with the mean of original variables was, however, low (.11). Also included was the category "pupils wait for turn" (II/7). On the basis of the placement of observer centroids on the discrimination dimension (see Figure 15), it was possible to establish that observers B and E deviated from the rest, most clearly from observer C and least from observer F, who, it will be remembered, represented a "wide coding of verbal communication" on the first dimension. Observers B and D tended more frequently than the others to use the categories "teacher initiates and terminates activity" (I/04), "teacher uses ideas, movement tasks suggested by pupils" (I/03) and "teacher participates silently in movement activity" (I/11). Observer C made exceptionally little use of these categories, but a frequent use of the categories "inter-pupil contacts free, range of ideas controlled" (II/2),

"pupils wait for their turn" (II/7), and "pupils' spontaneous activity" (II/4). In general, observer C used a more reduced method of coding a combination of verbal and non-verbal communication than observer E. It would seem that combining verbal and non-verbal communication, which is the central feature of this classification system, requires special alertness and a certain attitude. At least half of the observers strived consciously to do so.

While the first three dimensions brought out differences in the coding of infrequent or rapidly occurring categories, confused situations and non-verbal communication, the situation was quite different with the last two dimensions. In them were distinguished coders who used frequently occurring categories in certain ways:

- Matter-of-fact-centered coding of teacher talk - versus other and silence-centered coding - versus other.

The difference between observers was not significant on the last two dimensions, even though it yielded a reasonable interpretation. It should be pointed out that, in general, the use of the most frequently occurring categories, such as I/05 and I/10 in ambiguous situations, is not recommended according to the instructions given in connection with this classification system or with the Flanders category system (see Rule 1, page 79-80). The discriminant analysis brought out this problem of reliability variability of coding. Also, the shift phenomenon was highlighted in the interpretation of the last two dimensions.

6.4.5 Discussion of variability of coding results

Structure of the observer group

The discriminant functions that describe independent factors causing disagreement among coders were interesting from the point of view of theory. Observers could be placed into a certain group, which reflected their coding behavior. These discriminant functions were found to be associated with certain kinds of situations such as teacher-pupil verbal communication-centered, disturbed pupil activity situations or nonverbal communication-centered situations. The structure of the discriminant model reflected different coding decisions made in these situations and concerning the choice between infrequently (a) versus frequently (b) occurring categories:

DF I		DF II		DF III	
(a)	(b)	(a)	(b)	(a)	(b)

The central objective of the classification system was the identification of the sequence of teacher-pupil verbal and non-verbal communication, as well as the discrimination between directiveness and non-directiveness of the teacher's interaction with pupils.

Naturally, it was more difficult to observe teacher activity in a noisy and confused situation, because audibility was bad. Such situations are not, however, very common in observation studies, but they should be taken into ac-

count in analyzing the reactions of different observers and in refining categories and coding instructions. The technical equipment and the methods used for voice recording obviously need to be examined more closely.

The structure of the group of six observers with a similar training background was quite heterogeneous when examined in the light of differences revealed in their individual manner of using the metalanguage of the classification system. Coder differences emerged clearly in three linear factor groups of different composition (see Table 29 and Figures 14 and 15). As is usual in discriminant analysis, the first linear function predictor of disagreement separated one group (observer F) from the rest, then the next one (D) from the rest and so on. Observer variability was great, especially on the first three dimensions and in the discrimination space defined by two discrimination dimensions at a time (Figure 16).

On the basis of the nature of coding decision differences it was possible to get a description of the problems of the variability of coding in connection with the "testing of the instruments". Roughly speaking, about half of the observers approached coding in a way considered valid in terms of the theory, which, however, in this context, often took place at the cost of reliability.

As Flanders (1967b) noted in considering the problems of observer training as reliability, that ground rules two and fours seem to be an invitation to biased observation. Yet there is a theory of the "unbiased, bised observer", which recognizes that the observer is biased in the sense that his categorization must be consistent with his general assessment of the teacher's intent for a given sequence of action, but he is unbiased in that he remains open to all evidence that the general intent of the teacher may be changing. During preliminary training, the problem of distinguishing these shifts in categories usually arises. The solution is never fixed or final, but "the observer must learn to be sceptical of verbal habits which are often unreliable cues compared with the total time the teacher talks, the nature of the learning activities, and other more general evidence" (Flanders, 1967b, 159). Multiple coding with category clusters is the most flexible system but standardizing the observation procedures and establishing observer reliability may prove difficult.

In this case at least, the heterogenous group of coders offered a good basis for the discrimination of systematic differences, the shift phenomenon and factors that reduced inter-coder agreement. Thus it can be noted that by using a team of observers in the study the universe of generalizability could be broadened. But, in which direction it should be broadened is a question that also merits consideration when the measuring instrument is being refined (McGaw et al. 1972).

About validity coders

Although the results of the discriminal analysis can only be regarded as tentative on account of the nature of measurement scale it yielded quite useful information for the development of the instrument. The empirical findings reported in this study established clearly that high frequencies of occurrence are not necessary prerequisites for the reliable measurement of behavior. Certainly, if a particular type of behavior is of sufficient interest, we should not be deterred from attempting to measure it solely on the grounds that its occurrence is relatively infrequent. Nor, on the other hand, can we assume that the accumulation of large numbers of observations of a particular type of behavior provides some kind of guarantee that we have achieved precision of measurement.

What really matters, then, is not the number of times that a particular type of behavior has been observed, but whether the subjects of the observation have differed consistently in the extent to which they display that behavior. This cannot be inferred from considerations of frequency alone, but need to be determined by an analysis of inter-coder agreement and between-coder agreement of the type described earlier or those reported by Bookhout (1967), Heinilä (1976) or Komulainen (1973).

The construct of discriminators found in this study describes patterns of teacher and pupil behavior, which in Bookhout's (1967) study were found to be related to the social emotional climate. The quantity of positive emotive expressions of teacher and pupil talk (DF I), and the sequence of verbal and non-verbal interaction (DF II) also distinguished situations where teacher and pupils were moving and teacher was participating in movement activity (DF III), causing disagreement among coders. Decisions concerning the level of different forms of pupils' collective activity, operationalized as social access, were also reflected in results describing variation between coders. Also the results obtained by Tavecchio (1977) and Splinter (1980) suggests similar difficulties in coding interaction processes in physical education classes objectively. "Distinction like 'collectively' versus 'individually' or 'divergent' versus 'convergent' are not easy to make" Splinter (1980, 76).

In the present study, the inverse character of reliability and validity was highlighted, which had already been noted by Flanders (1960, 1967b 158-166, 1970) in his analysis concerning the training of observers and reliability problems.

6.5 Phase III: Investigation of the construct validity and sensitivity of the observation instrument PEIAC/LH-75

This section of the dissertation will report on an investigation of the construct validity and sensitivity of the observation instrument PEIAC/LH-75 using a cumulative multivariate analysis of the factorial structure of instructional situations, a grouping analysis based on the factor scores and test of discriminating power of categories.

The general principles underlying factor analysis and its various phases are well known. Only some special problems will be considered in this connection, after which the specific areas of multiple grouping analyses will be discussed.

A great deal of correlational research on validity employs factor analysis, which reorganizes a table of correlations to emphasize convergence. Reducing the central core of this information to a compact table of factor loadings often has a clarifying effect (cf. Cronbach 1971, Medley 1982).

In this connection, an attempt will be made (a) to use factor analysis as a means of reducing the dimensionality of the set of three cluster variables by taking advantage of their intercorrelations, and (b) to find ways of identifying fundamentally meaningful dimensions of the multivariate construct under study. This kind of evaluative research may be termed a method of controlled correlation to highlight the central roles of correlation coefficients as a primary index of the strength of relation, explanation, or prediction. Regarding kinds of possible conclusions, they will be probablistic in nature, reducing uncertainty, but not completely eliminating it (cf. Cooley & Lohnes 1976).

In this study the greatest interest centers on correlations between the original variables and factors. The matrix of scores of the categories of the three-cluster correlations forms a factor structure. This matrix will be used here primarily as an interpretative device, just as it is in any multivariate analysis which results in a factoring of a measurement battery. Here the same factor matrix is regarded as expressing both the theoretical composition of a measurement, thus "explaining" the measurement, and the correlations of the factor with the measurement "explaining" the factor (Cooley & Lohnes 1971).

When working with ipsative nominal scales, it is necessary to interpret the two poles of each factor separately. This situation is in general attributable to the use of taxonomies. "As the system is always in some state, an increase in any one form of behavior leads to a decrease in the other forms" (Komulainen 1971a, 16).

By using a three-cluster category system the variables are tied to ipsativity in more than one-way: between the categories within each cluster and between the categories of different clusters. Thus, we can discuss inter-cluster ipsativity and between-cluster ipsativity. A factor analysis will be employed in this context as a means of exploring ipsativeness on the construct under study.

The set of data analyzed here are the same as used before: data recorded on videotape during the autumn term of 1973. The data were gathered by six trained observers coding each situation three times: first in the live classroom

and then twice more with the videotape at one month intervals. The data set includes 24 P.E. lessons with a total of 28,800 six-second time units.(Table 5).

6.5.1 Aims of the factor analysis

This analysis will explore, from the point of view of the validity of Flanders' theory, the interaction in 24 P.E. lessons by considering the systematic variance among scores when using the PEIAC/LH-7S three-cluster category system on the construct under investigation.
In this phase of the study, the aims were:
1. to examine interaction in physical education classes by means of the factor analytical r-technique
- to identify the structural dimensions of interaction,
- to consider whether they correspond to logical dimensions or to the theoretical framework, and
- to consider the behavior of the emerging factors (factor scores) in combination with certain other variables (frame factors) as classified in accordance with the sex of the teacher, grade level and physical education subject area,
2. to explore the formation of homogenous groups of lessons in grouping analysis based on factor scores, and
3. to explore the formation of the factors predicting variability and grouping of lessons, "known" to be different.

6.5.2 Procedures

Selection of variables

The establishment of a minimum acceptable reliability for variables to be submitted to factor analysis was based on the following principles: Since there were no previous studies using this observation instrument, reliability of the data could not be presumed. The higher the reliability cut-off point set, the fewer variables would be submitted, and the greater the risk of throwing away valuable data. On the other hand, the lower the cut-off point, the greater the risk of diluting the factor analysis with so much worthless data that a great many poorly defined factors would be required to account for total variance. For this quasi experiment, the intention was to submit to factor analysis those variables, which might contribute significant loadings to factors. Estimating reliability by using the Kendall coefficient of concordance (*W*), 23 of the 27 categories were significant at the 0.01 level. The remaining four were categories with low frequencies and/or indicating a confused situation (I/03, I/12, II/8, III/7). In the light of this criterion, a total of 27 variables were submitted for analysis. The results reported here are based on a video-recorded observation (T_2), in which the level of reliability was the highest of the three rating times. The means of Scott's *pi*, computed from the scores of the six trained observers, were, by clusters, .61, .71 and .77.

Factoring and principles of interpretation

The intercorrelation matrix was obtained by correlating the three-cluster category frequencies 27 x 27 computed from the six observers' scores (total 28,800 six-second time units) in the lessons (n=24). The data from three coding occasions (Appendix C.I) were subjected to factoring separately. The correlation matrices were factored by using the principal axis method, and the numerically highest correlations were used as estimates of h^2. Rotation was carried out by the varimax technique. This rotation method was chosen because, being orthogonal, it was likely to yield a simple and clear-cut result useful at the initial stage of this "structure seeking" investigation.

The number of factors to be rotated was determined according to the principle that (1) it is preferable to include too many than too few factors, and (2) a description that is optimal both interpretationally and in terms of the simple structure rule should be sought with successive reductions of the primary base. Four, five, six, seven and eight factors were rotated with the varimax technique.

Seven factors proved to be the most interpretable and stable combination. The consistency of the structures of the seven-factor varimax resolution was examined by analysing the factor structure computed from three data sets (coding occasions T_1, T_2, T_3) by means of Symmetric Transformation Analysis (Appendix 3.2). Each factor extracted was interpreted as a structural dimension by studying the categories with appreciable loadings (-.30), synthesizing them, and naming the composite pattern.

The factor loadings of categories and the regression coefficients by them (Appendix 3.3) in the estimation of factor scores helped to identify the categories that were central in the construct of the factor in question. In addition, the lessons for which the factor scores were the highest were compared with those with the lowest factor scores.

6.5.3 Results of the factor analysis

Correlations between categories of the three clusters

The correlation matrices between categories of different clusters (Table 30) express the interdependence of the categories of each cluster throughout the lessons observed. The figures are in general so low that categories may be considered sufficiently independent of each other to meet the requirements of independence imposed on observational methods. Using ipsative nominal scales, it is evident that there will be some high negative correlations, and as stated before, the process is always in some state. Therefore an increase in any one form of behavior leads necessarily to a decrease in the other forms. For instance, in the verbal cluster (I) the category indicating teacher's silent behavior (I/10) and the category indicating the most dominant teacher's verbal behavior (I/05) correlated negatively. Also it is understandable that there will be positive correlations between the categories of initiative behaviors and response behaviors. Categories of different clusters correlated with each other

both positively and negatively. The highest positive correlation, .98-.97, was found between categories II/8 and III/7 of clusters two and three, both indicating a confused situation. These categories were always used together in beginning and finishing coding.

TABLE 30 Categories of the three clusters on correlation matrix for observation T_2.

Variables cluster categ. No																											
I	01																										
	02	55																									
	03	-02	05																								
	04	-17	-35	-14																							
	05	10	32	-02	-19																						
	06	02	-10	09	54	-02																					
	07	-00	-02	32	06	29	03																				
	08	-04	-18	51	45	20	31	52																			
	09	-26	18	50	-23	48	05	43	49																		
	10	-24	-19	-09	-31	-60	-47	-16	-36	-35																	
	11	-15	-37	-15	01	-31	-08	-34	-26	-30	-21																
	12	00	-08	-01	09	-02	12	04	06	-05	-04	-00															
II	1	-27	-25	-21	54	-11	72	-21	09	-22	-24	03	30														
	2	46	33	13	-44	-17	-51	07	-23	14	54	-38	-19	-65													
	3	-21	-26	-06	-12	-25	-10	-04	-17	-28	-11	77	-08	-18	-41												
	4	06	22	17	-10	07	-12	04	-06	54	-14	06	05	-19	13	-03											
	5	-20	05	09	41	61	25	03	46	08	-57	-18	01	26	-55	-20	-16										
	6	-13	-02	04	-19	58	-14	38	31	58	-08	-45	08	-17	-02	-31	19	25									
	7	-44	-39	-08	14	-07	17	08	09	-19	-25	56	21	19	-57	59	-10	07	-06								
	8	21	-05	15	13	17	-14	50	58	47	-19	-12	05	-13	01	-15	19	24	26	-02							
III	1	07	07	15	36	-14	47	-12	-24	03	-32	18	-18	41	-16	-03	15	-08	-34	-08	-23						
	2	-39	-42	-18	18	21	-12	13	32	08	11	-20	39	18	-20	-19	-21	34	27	07	22	-59					
	3	41	66	06	-36	22	-10	04	02	20	-07	-27	-19	-41	34	-18	10	06	18	-17	-11	-19	-47				
	4	30	-06	-23	-23	-35	-21	-10	05	-10	38	03	-08	-12	25	08	-10	-38	-16	-15	25	-26	-06	-18			
	5	-33	-38	19	-20	-29	-22	-09	-24	-10	11	62	-12	-25	-15	76	06	-35	-20	48	-13	13	-24	-31	-08		
	6	-17	03	51	40	-07	01	19	42	-06	04	-15	-08	-15	-02	04	-07	22	-00	16	-11	-08	-05	22	-13	-10	
	7	10	28	12	-11	10	-07	-00	-05	52	-16	00	07	-15	12	-10	98	-10	18	-15	17	09	-18	15	-05	-08	-05
		01	02	03	04	05	06	07	08	09	10	11	12	1	2	3	4	5	6	7	8	1	2	3	4	5	6
		I												II								III					

Results

Factor analysis yielded seven factors accounting for 68.6% of the total variance (Table 31). Factor scores were estimated for every lesson in the seven factors. The results are illustrated in Figure 16a-g, based on the means and dispersions of factor scores and demonstrating the location of each lesson in structural dimensions as classified according to the sex of the teacher, the grade level, and the physical education subject area.

It was found that the positive pole activities consisted mostly of the teacher's verbal activities. However, in the first factor a type of non-verbal form of teacher activity, participation in student activity (I/11), was evident (Figure 16h-g). The teacher's silent behavior as guidance (I/10), which is a common type of activity in ball games, was characteristic of the negative-pole activities. Two factors, IV and V, had high loadings only in the positive pole.

TABLE 31 Varimax-rotated factor matrix

Cluster	Cat.	1	2	3	4	5	6	7	h^2
I	01.	29	-03	-64	28	01	-13	-32	68
	02.	37	11	-66	-10	16	24	-12	69
	03.	-01	-07	-13	35	18	-06	65	60
	04.	04	-68	29	11	-13	02	29	67
	05.	18	14	-09	23	08	79	-11	75
	06.	06	-82	01	04	-05	16	05	70
	07.	01	12	04	61	03	12	27	47
	08.	12	-20	16	72	-10	23	36	79
	09.	17	10	-17	58	51	25	10	73
	10.	20	48	23	-29	-11	-62	08	81
	11.	-84	-13	00	16	03	-13	-21	81
	12.	06	-10	36	05	06	07	-15	17
II	1.	09	-76	40	18	-10	10	-17	82
	2.	44	51	-37	02	10	-47	05	82
	3.	-90	08	-05	-06	-09	-06	-02	83
	4.	-03	04	-06	08	98	00	-00	97
	5.	13	-25	16	16	-17	71	12	68
	6.	25	35	21	30	21	47	06	58
	7.	-67	-11	26	-00	-11	21	08	59
	8.	07	05	10	78	10	02	-13	65
III	1.	-10	-66	-29	-22	-22	-13	-07	64
	2.	18	16	75	17	-20	20	-09	74
	3.	25	28	-64	-10	04	30	15	67
	4.	05	17	-03	19	-14	-49	-33	44
	5.	-82	18	03	-06	09	-23	16	79
	6.	03	-04	-05	06	-13	-08	71	53
	7.	08	02	-07	05	95	05	-05	92
Eigenvalue		3.4	3.1	2.7	2.5	2.5	2.7	1.7	18.5
%		12.5	11.5	10.0	9.3	9.2	10.0	6.2	68.6

The factors obtained are shown below. The first factor was clear-cut in content. Here, all the most important loadings were negative. The loadings were spuriously high. The social access and the social form cluster categories (Cluster II and Cluster III) had high loadings on this structural dimension as well as teacher's silent participation in movement activities in situations where inter-pupil contacts were free and range of ideas open, work divided among groups or individuals. The positive pole activities consisted of the teacher's verbal positive reactions and corrective feedback to the pupils' activities. Comparing the different lessons by considering the factor scores estimated for them, the lesson of rhythmic movement-expression showed the highest loadings in this factor. These variables are descriptive of the entire indirect influence area. This structural dimension was labelled "indirect nonverbal integrative idea generation -- teacher's verbal communication and motivation".

Factor I:	Cluster/ Category	Positive pole	Cluster/ Category	Negative pole
	I/02	+.37	II/3	-.90
	II/2	+.44	I/11	-.84
			III/5	-.82
			II/7	-.67

The content of Factor II also was clear. The negative pole concerned the teacher's verbal direct communication and its intensity in the situation in which inter-pupil contacts and movement activities were restricted and range of ideas controlled. The positive pole was associated with situations in which the teacher's silent guidance was predominant and in which inter-pupil contacts were free but the range of ideas was still controlled. All the woman teacher's gymnastic lessons showed high loadings on this factor (Figure 16b, p. 174). The structural dimension was descriptive of the entire direct influence area. It was labelled "teacher's total, intensive guidance (+)/teacher supervision and organization (-)".

Factor II:	Cluster/ Category	Positive pole	Cluster/ Category	Negative pole
	II/2	+.51	I/06	-.82
	I/10	+.48	II/1	-.76
	II/6	+.35	I/04	-.68

Factor III consisted of categories from all three clusters. In the positive pole the highest loading was related to situations where the class was divided by uniform task, and the second highest variable loading described the social situation in which inter-pupil contacts and movement activities were restricted and the range of ideas controlled. The dominating characteristics of the negative pole were the teacher's positive verbal reactions to pupil activities, specificity of supportive supervision in the situation in which the class was divided, the tasks differentiated, and the range of ideas controlled. In this factor, the apparatus and gymnastics lessons, especially of the male teacher, showed high loadings (Figure 16c). This structural dimension was labelled "specifity-uniformity of teacher's nonverbal guidance: -specificity of verbal supportive supervision uniformity of teacher guidance". These aspects are descriptive of the entire direct/indirect influence area.

Factor III:	Cluster/ Category	Positive pole	Cluster/ Category	Negative pole
	III/2	+.75	I/02	-.66
	II/1	+.40	I/01	-.64
			III/3	-.37
			II/2	-.37

In Factor IV, all the most important loadings were positive. The fourth factor was related to confused situations where the dominant characteristic was pupil-teacher verbal communication, which consisted particularly of pupils' suggestive activity. The dimension was typified by the high loading of teacher's acceptance of pupils' spontaneous activity as well as by the loading of teacher's criticism. One low level gymnastic lesson in particular had high loadings on this factor (Figure 16d). This dimension was labelled "directing communication".

Factor IV:	Cluster/ Category	Positive pole	Negative pole
	II/8	+.78	-
	I/01	+.72	
	I/07	+.61	
	I/09	+.58	
	I/03	+.35	
	I/06	+.30	

Factor V was typified as non-structured situations in which the social form as well as social access was unclear. In this context pupils were asking for instructions and expressing their own ideas. Only one low-level apparatus lesson had exceptionally high loadings of this factor (Figure 16e). The dimension was labelled "non-structured spontaneous Pupil activity".

Factor V:	Cluster/ Category	Positive pole	Negative pole
	II/4	+.98	-
	III/7	+.95	
	I/09	+.51	

The dominant characteristics of the sixth structural dimension were phases of the lesson as orientation and work typified by verbal/nonverbal interaction. The positive pole mainly concerned the teacher's presentation of information, pupils following instructions, organizing themselves and assisting in organization. The negative pole was associated with activity situations in which the class was divided, tasks distributed among groups and within groups, the range of ideas controlled and silent guidance predominated. The female teacher's apparatus and rhythmic movement-expression lessons had high loadings on this factor (Figure 16F). The structural dimension can be named "teacher-dominant verbal subject centricity - non-verbal group work centricity".

Factor VI:	Cluster/ Category	Positive pole	Cluster/ Category	Negative pole
	I/05	+.79	I/10	-.62
	II/5	+.71	III/4	-.49
	II/6	+.47	II/2	-.47
	III/3			

Factor VII was typified by the teacher's verbal response behavior. The positive pole was related to situations in which pupils worked individually, tasks were differentiated and the teacher stimulated the pupils, activity and thinking by acceptance of their movement ideas.

The negative pole was related to situations in which the class was divided, tasks were distributed among groups and within groups, and the teacher encouraged different groups by acceptance and praise. The rhythmic movement-expression lessons of both teachers had a high loading for this factor (Figure 16g). This factor was labelled "Attributing teacher's response behavior to individuals/groups".

Factor VII:	Cluster/ Category	Positive pole	Cluster/ Category	Negative pole
	III/6	+.71	III/4	-.33
	I/03	+.65	I/01	-.32

The factor structure by frame factors

The behavior of the resultant factors was considered in combination with certain variables, and frame factors, as classified according to the sex of the teacher, grade level and physical education subject area. The results are illustrated in Figures 16a-g and 17.

For the factor scores reported for the two teachers in Table 32, a high Factor I score indicates a predominance of behaviors extending the pupil's freedom of action, whereas a high Factor II score indicates an accentuated part played by teacher initiation and direct communication, reducing the pupil's freedom of action. Factor III indicates a uniformity of teacher guidance and specificity of silent guidance in situations such as ball games and apparatus work.

TABLE 32 Significance of differences between factor scores estimated for the two teachers (man-woman) (24 lessons, n=12) (ANOVA)

	Man Teacher		Woman Teacher		df.=22
Factor	X	SD	X	SD	t
1	500	104	501	102	.00
2	465	48	536	428	-1.82
3	524	83	476	116	1.19
4	532	121	468	72	-1.57
5	519	141	481	22	-.94
6	467	114	533	80	1.65
7	477	82	523	118	1.11

There were found no statistically significant differences between the two teachers.

The differences in teaching in relation to the three grade levels (Table 33) were clearest in Factor IV. The amount of directing communication varied according to the age of the pupils.

TABLE 33 Significance of the difference between factor scores estimated for the lessons of three grade levels (24 lessons, n=8) (ANOVA)

	Low Grade		Middle Grade		Upper Grade		Low-Middle df=14	Low-Upper df=14	Middle-Upper df=14	df=20
Factor	X	SD	X	SD	X	SD	t	t	t	F
1	482	21	498	101	521	148	-.43	-.73	-.38	.29
2	509	90	507	123	485	101	.03	.49	.38	.12
3	473	86	512	111	515	115	-.77	.83	-.07	.40
4	570	139	479	70	451	39	-1.66	-2.32	-.96	3.59
5	539	172	490	28	471	9	-.79	-1.11	-1.84	.96
6	54	114	476	71	475	111	-1.53	-1.30	.01	1.40
7	554	150	486	56	460	55	-1.20	-1.67	-.97	1.99

variance is not equal between groups, - = $p < .05$

The differences between subject areas in relation to the factor structures (Table 34) were great and clearest in the first three factors. Rhythmic movement-expression differed from the others in the first dimension; gymnastics in the second dimension, and apparatus and ball games differed greatly from each other in the third structural dimension. In this context, gymnastics and apparatus were similar to each other and differed from both ball games and rhythmic movement-expression. On the other hand, in the lessons of ball games and rhythmic movement-expression, the interaction was uniquely almost silent, differing from the communication of the other subject areas.

TABLE 34 Significance of differences between factor scores estimated for the four subject areas ANOVA

	Subject	Gymnastics N=6		Apparatus N=6		Rhythmic N=6		Ball games N=6		1-2 df=10	1-3 df=10	1-4 df=10	2-3 df=10	2-4 df=10	3-4 df=10	df=3 df=20
Factor no	area	m	d	m	d	m	d	m	d	t	t	t	t	t	t	F
1		475	14	450	13	660	115	435	31	3.21	-3.49	2.91	-4.33	1.09	4.22	14.88
2		631	118	451	26	476	48	442	45	3.65	2.98	3.67	-1.14	.44	1.29	9.97
3		506	66	611	63	499	48	385	79	-2.82	.21	2.88	3.47	5.49	3.01	12.12
4		553	191	481	37	492	37	475	68	-.91	.77	-.95	.53	-.19	-.54	.71
5		478	17	547	203	497	32	478	9	.83	1.31	.16	-.59	-.82	-1.34	.59
6		497	45	563	82	474	108	466	143	1.71	-.50	-.51	1.61	1.43	-.10	1.12
7		469	54	477	37	585	170	469	54	.29	1.60	.01	1.53	-.27	-1.59	2.15

$p < 0.01$ N = 24 lessons; the mean = 500; standard deviation = 100; 6 observers; 4800 six sec. time units; tot. 28800 time units

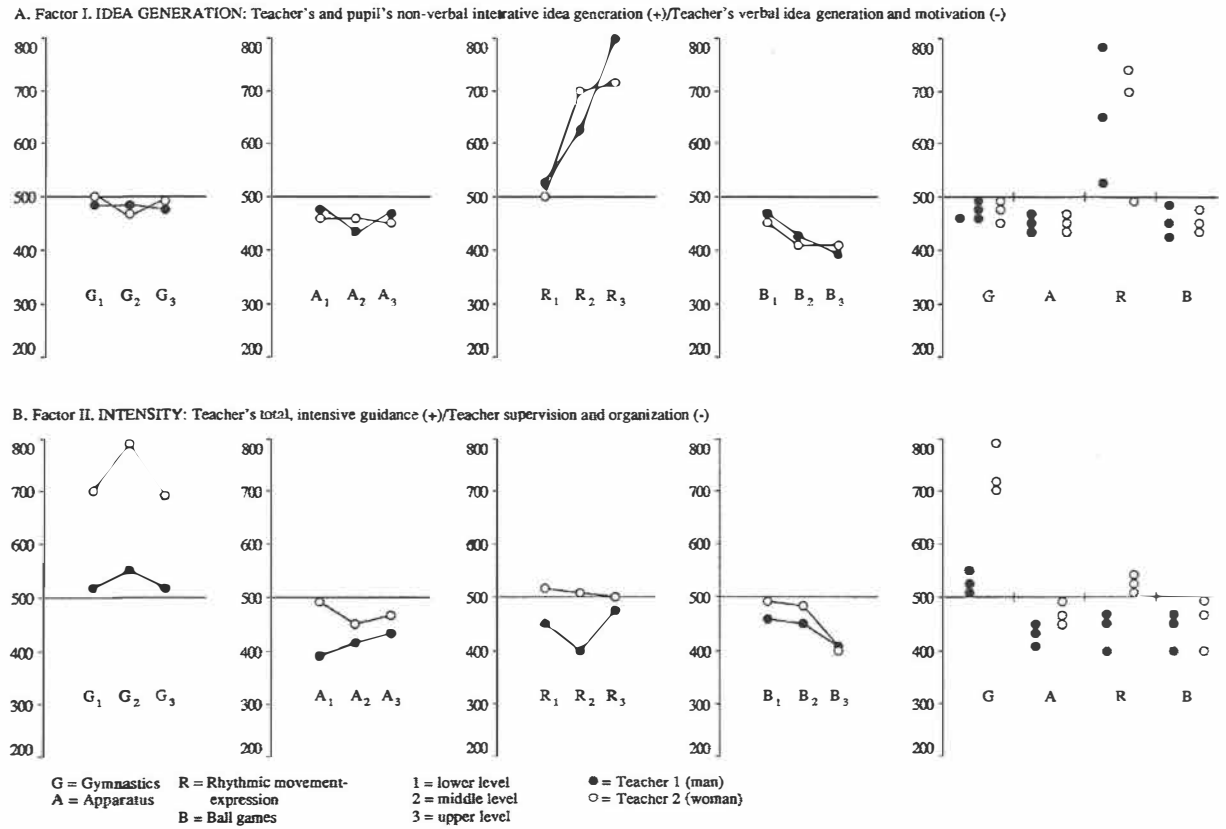
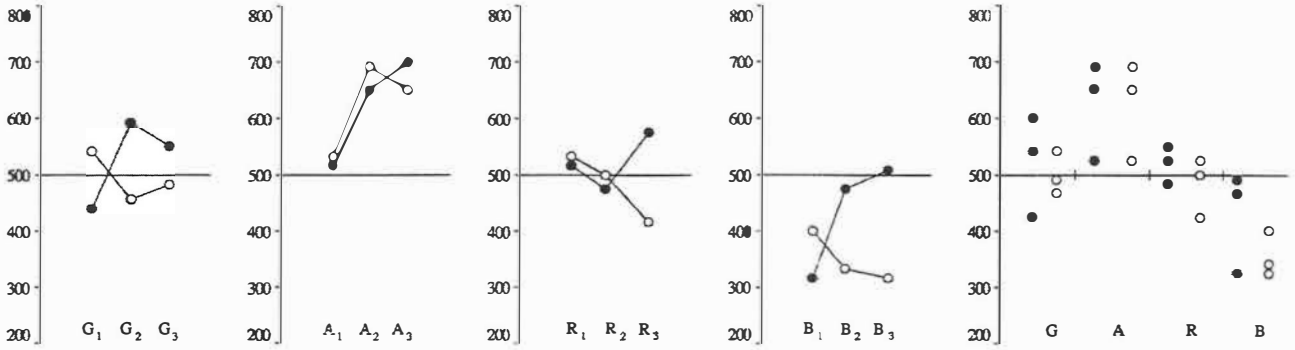
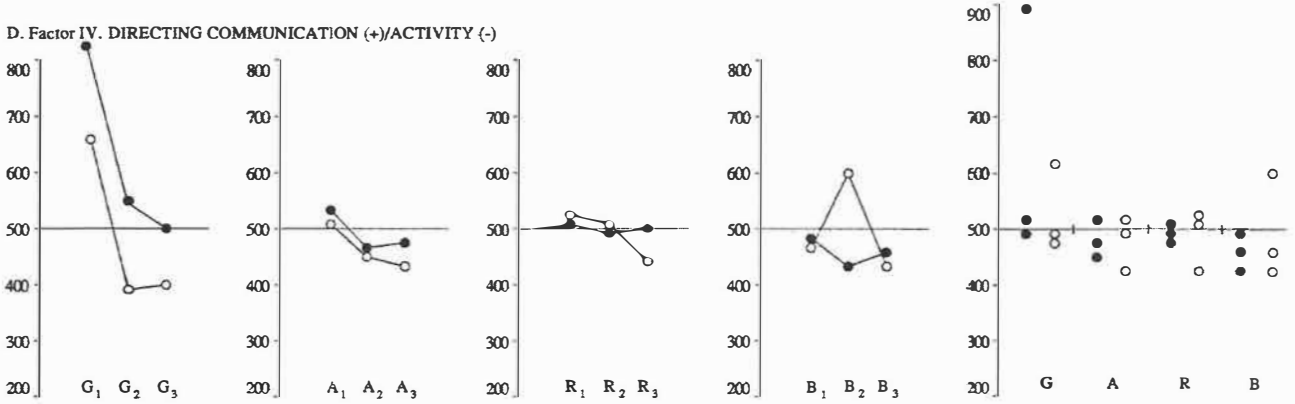


FIGURE 16 Location of each lesson in structural dimensions based on the means and dispersion of factor scores I – VII (continues)

C. Factor III. SPECIFICITY-UNIFORMITY OF QUIDANCE: Specificity of supportive supervision (+)/Uniformity of teacher guidance (-)



D. Factor IV. DIRECTING COMMUNICATION (+)/ACTIVITY (-)



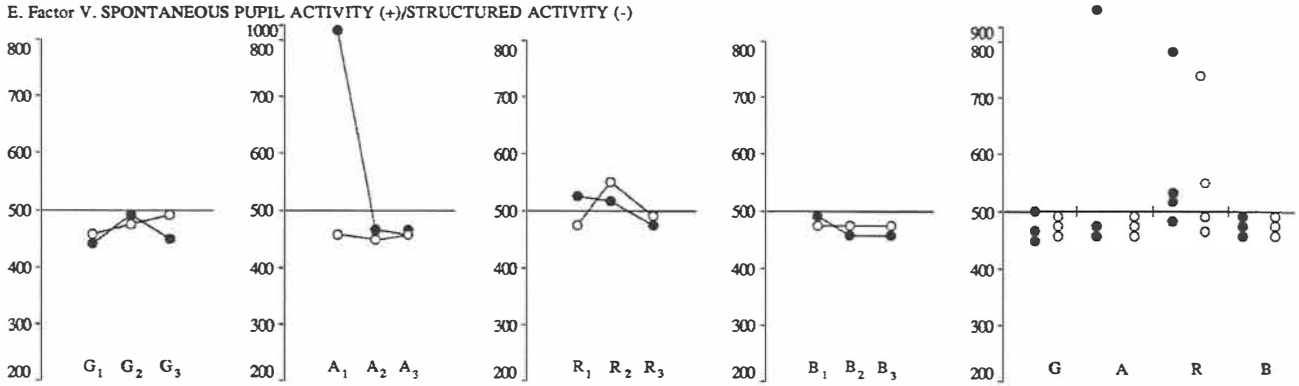
G = Gymnastics
A = Apparatus
R = Rhythmic movement-expression
B = Ball games

1 = lower level
2 = middle level
3 = upper level

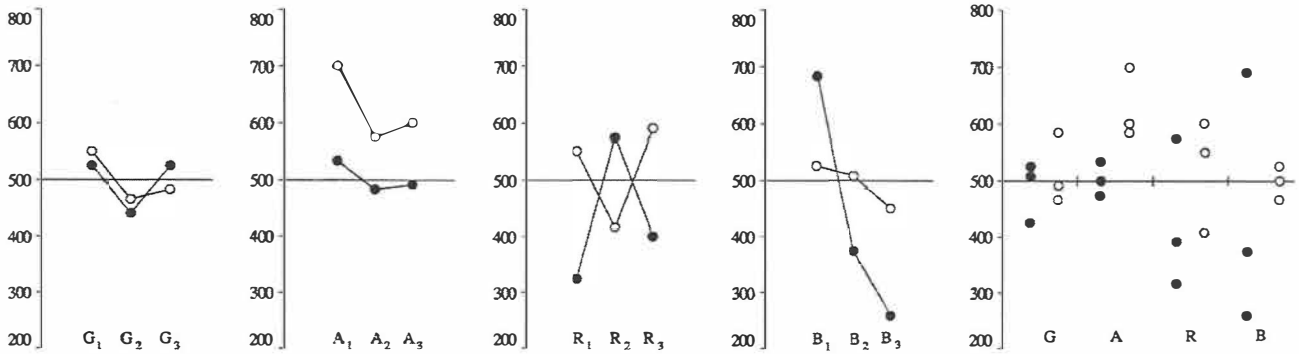
● = Teacher 1 (man)
○ = Teacher 2 (woman)

FIGURE 16 (continues)

E. Factor V. SPONTANEOUS PUPIL ACTIVITY (+)/STRUCTURED ACTIVITY (-)



F. Factor VI. SUBJECT CENTRICITY - PROCESS CENTRICITY: Teacher-dominant subject centricity (+)/Group activity centricity (-)



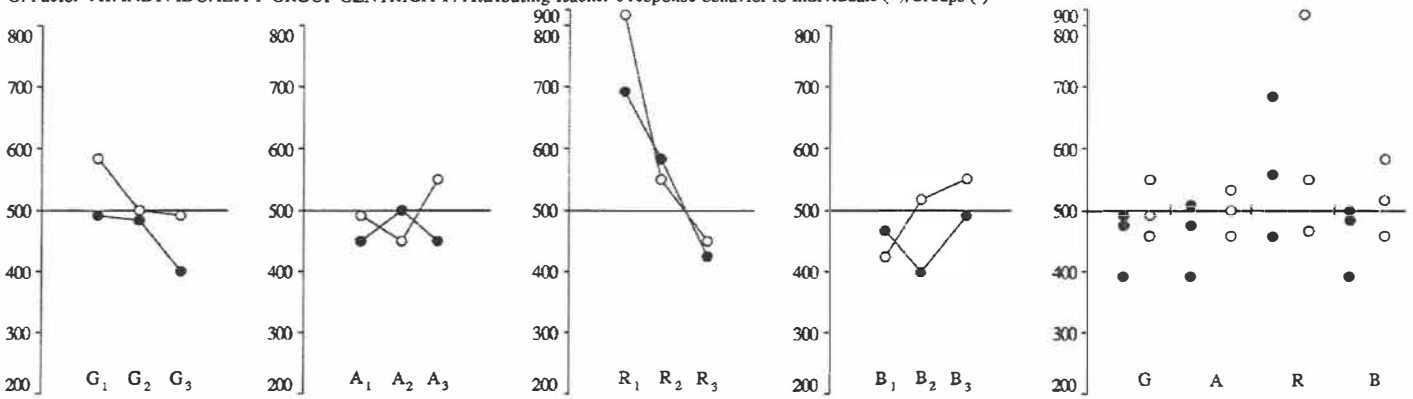
G = Gymnastics
A = Apparatus
R = Rhythmic movement-expression
B = Ball games

1 = lower level
2 = middle level
3 = upper level

● = Teacher 1 (man)
○ = Teacher 2 (woman)

FIGURE 16 (continues)

G. Factor VII. INDIVIDUALITY-GROUP CENTRICITY: Attributing teacher's response behavior to individuals (+)/Groups (-)



G = Gymnastics
A = Apparatus
R = Rhythmic movement-expression
B = Ball games

1 = lower level
2 = middle level
3 = upper level

● = Teacher 1 (man)
○ = Teacher 2 (woman)

FIGURE 16 (continues)

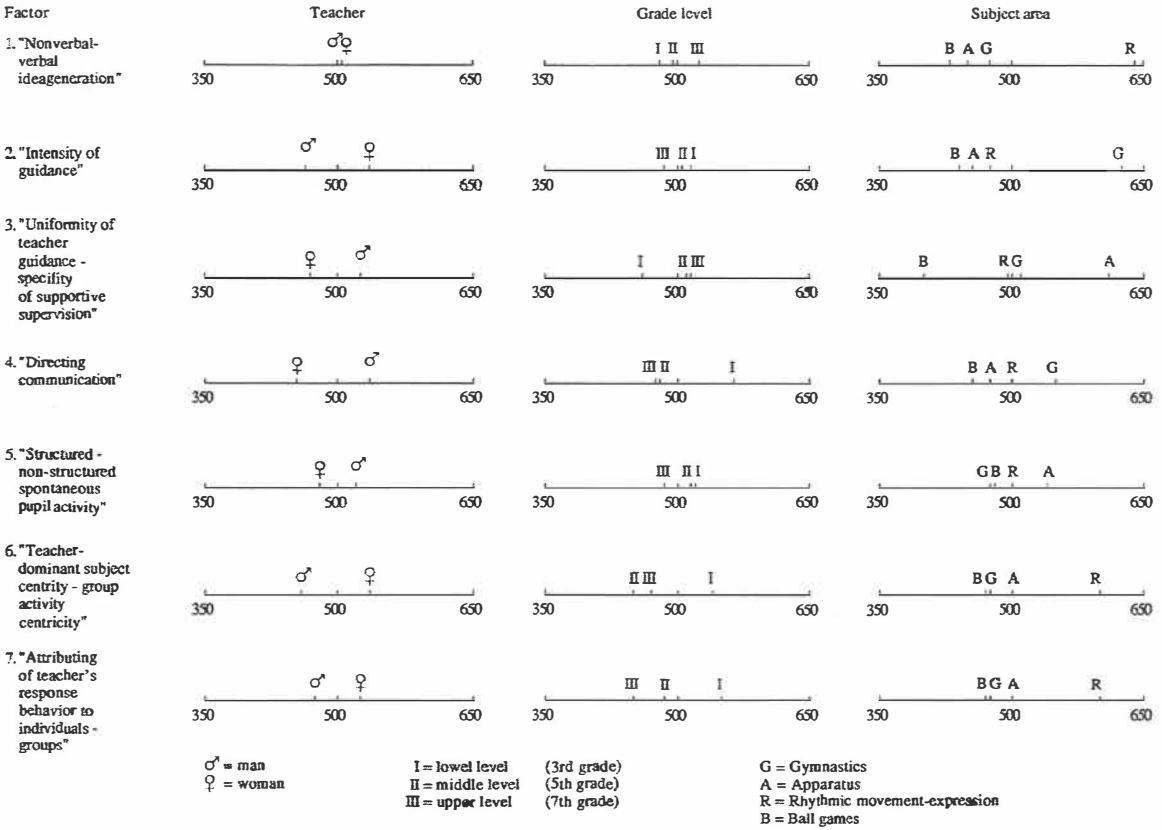


FIGURE 17 Summary: average location of different frame groups (teacher, grade level, subject area) in factor structure dimensions of physical education interaction, process (7 factors, Varimax solution)

6.5.4 Grouping analysis based on factor scores

Procedures used in grouping analysis

In the preceding section, the factor scores estimated for the lessons were considered by interpreting the content of the structural dimensions by comparing the factor scores and location of the lessons in the different dimensions when classified according to the sex of the teacher, grade level, and the physical education subject area. In this section, the significance of these frame factors will be determined by considering the results of grouping analysis based on the factor analysis.

In grouping analysis, the goal is to form groups for each of which the sum of distances from the group mean of observation will be minimum. The number of groups must be decided in advance. For this purpose 4 to 6 groups were formed because the factor analysis had yielded seven factors. All the HYLPGA groupings were repeated with three different initial values. The emerging groupings varied to some extent, depending on the initial values.

Results of grouping analysis and frame factor specificity

The results of the grouping analysis are presented in Table 35 and Figure 18, which illustrate the average location of the six lesson groups (1-6) on the seven-varimax factor dimensions on the basis of their means and standard deviations. The principal lessons of the factors were identified by considering both the results of the grouping analysis and factor scores.

TABLE 35 Estimated factor scores of the six groups formed by means of grouping analysis

Varimax	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6	
Factor	X	SD	X	SD	X	SD	X	SD	X	SD	X	SD
I	411	30	465	00	454	14	707	55	477	17	505	17
II	452	46	490	00	470	45	470	46	735	37	512	35
III	390	68	531	00	611	50	487	50	491	38	528	3
IV	534	157	527	00	483	31	478	34	467	117	518	12
V	476	11	961	00	469	13	500	30	483	10	491	26
VI	475	123	529	00	544	83	492	89	499	38	437	106
VII	471	46	473	00	462	48	489	60	497	31	777	99

It was found that the lesson groups were located at the positive pole in four of the seven structural factor dimensions and at both poles in Factor III. Thus, the behavior in these lesson groups was "known" characterised by the dominating features of these poles.

By considering the behaviors of the resultant factors and lesson groups in combination with the top factors, five factors appear to be connected with the grouping of lessons, and both poles of Factor III showed the most predictive power in the grouping of lessons (Table 36).

TABLE 36 Variation of six groups through principal factor, teacher, grade level and subject area.

Group no	Lesson no	Principal Factor no	Teacher Male	Female	Grade level Low	Middle	High	Subject area Gymn.	Appr	Rhytmic.	Ball
1	5,8,11,12 13,16,23	3(-) unif.	4	3	3	2	2	1			6
2	21	5(+) spont.	1			1			1		
3	1,2,7,18 20,22,24	3(+) specif.	4	3	1	3	3	2	5		
4	6,10,14, 15	1(+) expr.	2	2		2	2			4	
5	3,4,9	2(+) intens.		3	1	1	1	3			
6	17,19	7(+) indiv.	1	1	2					2	

Principal factor in grouping analysis

I Exressivity (4)
II Intensity of guidance
III Uniformity (1) specifity of guidance (3)
IV Directing communication
V Spontaneous pupil activity (2)
VI Subject centricity - nonverbal group work centricity
VII Individuality – group centricity non directive communication (6)

By considering the behavior of lesson groups in combination with frame factors, as classified according to the sex of the teacher, grade level and physical education subject area, it was found that there were two principal sources of variance in the set of lessons: the P.E. subject area, and the teacher. A possible third source of variance consisted of the interaction between the first two, and a fourth, of the interaction between the first two and the grade level.

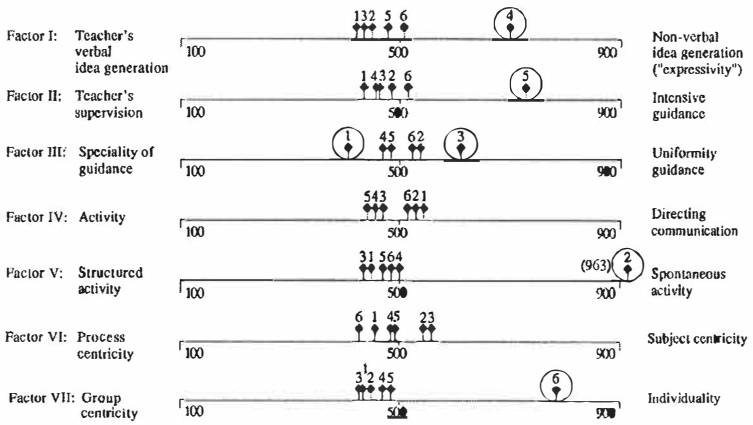


FIGURE 18 The average location of lesson groups' 1-6 on the varimax factor dimensions based on their means and standard deviations

Both in the factor analysis and in the grouping analysis, the lessons had a certain tendency to cluster according to the P.E. subject area. The teachers appeared to follow the traditional ways of teaching different P.E. subject areas. Or perhaps it was the subject area itself, its structure and content, that caused the teacher to choose a certain way of teaching, using direct or indirect

influence. Or maybe the measuring instrument was itself sensitive in describing this kind of behavioral differences. In any case, such grouping is regarded as too narrow (cf. Flanders 1965). Applying the concepts used by Cheffers (1973) we ask now: (1) Is the instrument sensitive enough to make the discrimination required for research problems (sensitivity), and (2) does the instrument possess the ability to distinguish between groups "known" to behave differently an the construct under study (construct validity)? A useful way to explore this question further is a cumulative evaluation of the results obtained in grouping analyses. In a pilot study this variance was examined more closely using discriminant analysis techniques in an effort to estimate the predictive power of categories of different clusters (Heinilä 1983).

Predictive power and sensitivity of the PEIAC/LH75

Because there was no external criterion available to assess the construct validity of this instrument, it was decided to use also multiple discriminant analysis for examining more closely the portion of variance through "criterion groups" which were predictable from or explained by the known variance on the linear combination of predictors (Cooley & Lohnes 1971). The design involved the assessment of two or more traits by two or more methods (See Heinilä 1983, 129).

Results of discriminant analysis

The data used were the score distributions of categories from the 24-lesson (T_2) as coded by six trained trained observers, and the 27 categories of the three-cluster classification system. The six lesson groups formed by using grouping analysis based on factor scores were structurally homogeneous and there were differences in the mean distributions of variables. (Appendix 5 and 6) (Heinilä 1983, 128-129).

The program selected 16 of the 27 classification categories and set them in sequence according to how much they increased the model's discriminating power. It is possible even on the basis of these categories to get an idea of the nature of the discrimination. The discrimination model included the seven categories of Cluster I (Verbal) four of Cluster II (Movement and Social Access), and four of Cluster III (Social Form). The categories of Cluster II, representing pupils' collective activity with the range of ideas closed and with open ideas, and the categories of Cluster III showed the most predictive power. Both categories, which occurred rarely and those occurring most frequently were represented in the model (Heinilä 1983).

From the structure coefficients of obtained and the nature of the factors, the five functions extracted appeared to measure:
DF I: Range of ideas for pupils, closed - open
DF II: The level of structuration: high - low
DF III: The level of intensity of guidance: high - low
DF IV: The level of specificity of non-directive guidance high - low

DF V: The media of non-directive communication (attributing teachers response behavior to individuals/groups): non-verbal - verbal

Thus, these discriminative dimensions describe different aspects and levels of "teacher's control of students' freedom of action", which is the feature that Flanders (1965) gives as the main purpose of interaction analysis. Teacher's influence is connected strongly with the content and way of communication in instructional process as well as with pupil's opportunity to social contacts.

On the basis of the nature of structural differences found in these analyses it was possible to describe the problems and level of the discriminant validity and sensitivity of the "testing of the instrument". The lesson groups' variability was large, especially in the first three dimensions, and in the discrimination space defined by the first and the other discriminant dimensions.

The structure of the discriminative model was congruent with the structure of the measuring instrument and produced the following sequence predicting the grouping of lessons, illustrated in Figure 19.

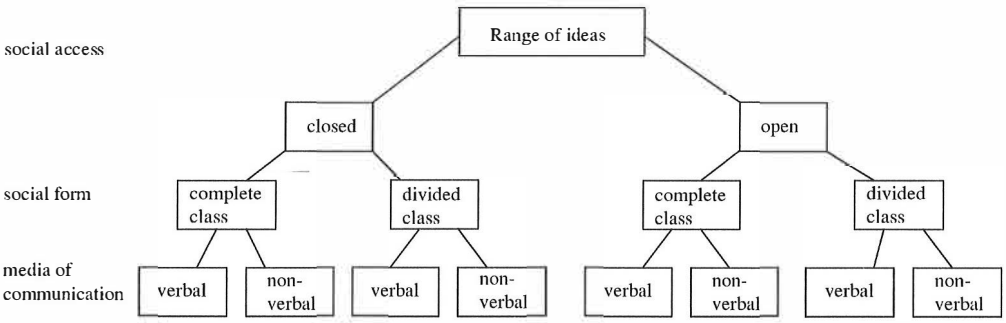


FIGURE 19 The components of the instructional process, the relationship between them and strategy used in connection with the application of the PEIAC/LH-75 (Heinilä 1983, 1987)

6.6 Summary, discussion of results and conclusions

The first phase explored interaction in 24 P.E. lessons by means of the factor analytical r-technique from the point of view of construct validity of Flanders' theory. The second phase examined the formation of homogenous groups of lessons in a grouping analysis based on the factor scores. The nature of the factors and lesson groups were considered in combination with the frame factors of the study.

It was found that there seemed to be two principal sources of variance in this set of lessons: (1) the P.E. subject area and (2) the teacher, and perhaps two others consisting of (3) the interaction between the first two, and (4) the interaction between the first two and the grade level. The principal lessons of the factors were identified by considering both the results of the grouping

analysis and the factor scores. It was found that the lesson groups were located at the positive pole of four of the seven factorial dimensions and at both poles of Factor III. Thus, the behavior in these lesson groups was "known", characterized by the dominating features of these poles.

In the third phase, an attempt was made to determine more closely the predictive power and sensitivity of the category system by using discriminant analysis technique (Heinilä 1983).

The program selected 16 of the 27 classification categories and set them in sequence according to how much they increased the model's discriminating power. The categories of Cluster II (Movement and Social Access) and the categories of Cluster III (Social Form) showed the most predictive power. Both the categories which occurred rarely and those, which occurred most frequently, were presented in the model:

DF I: Range of ideas for pupils: open - closed
DF II: The level of structuration: high - low
DF III: The level of intensity of guidance: high - low
DF IV: The level of specifity of non-directive guidance: nonverbal - verbal
DF V: The media of non-directive communication (attributing teacher response behavior to individuals/groups): non-verbal - verbal.

Although the results of these multiple analysis can only be regarded as tentative on account of the nature of the level of the measurement scale, they yielded quite useful information for estimating the construct validity and sensitivity for the development of the instrument.

Lessons could thus be placed in a certain group, which reflected their aspects of direct-nondirect teaching in a non-verbal and verbal context. The quality of teachers' verbal behavior had more prediction power in the grouping of lessons than the quantity of it. The quantitity and quality of teachers' nonverbal behavior posed a high predictive power in the classification of lessons. The principal sources of variance in the classes observed appear to be the subject area and the teacher, and to a lesser degree the interactions among the three frame factors. The variables of the clusters identified as "movement and social access (II) and social form (III) showed the most predictive power of the category system. The contextual variables, as subject matter were related to them.

The results obtained in analysis describing factors that predicted the grouping of lessons among criterion groups were interesting from the point of view of theory (Flanders 1965, 18, 1970). The classification of activities presented by Flanders (1965) is based on purpose of particular activity: planning, work, evaluation, and administration. These units are further subdivited according to whether the teacher or pupils are initiatiors for that activity (Heinilä 1971). In connection with Flanders Interaction Analysis Category System (FIAC), Flanders (1970) used the concepts: "pupils' perception of the goals", "teacher influence or initiative", "pupils opportunity to social contacts", and "freedom of ideas" (see p. 35-36). Even if temporal units within the flow of teaching process are not used, it can be considered in agreement with (Koskenniemi 1981, 44) that the Flanders' classification system is based on the "pupil's intentionality". As an adaptation of Flanders' Interaction Analysis

System the PEIAC/LH-75 operates with the same concepts: "pupil's perception of the goals", "teachers influence or initiative" (Cluster I) pupil's *"freedom of ideas and activity" (Cluster II) "pupils opportunity to social contacts"* (Cluster III) (see p. 80-82). By using PEIAC/LH-75 also the flow of teaching process is recorded by using 6-second time unit and tripple coding is made to categories of three clusters and the codings are analysed in matrixes.

Based on results of these multiple analysis it was evident that the classification of interaction process in physical education classes is also based mostly on "pupil's intentionality", and "opportunity to social contacts" as connected with the content and form of learning (cognitive, affective, and psychomotor proprieties of the verbal and non-verbal communication as well as pupil's movement activity). Thus, content cannot be regarded only as something offered to students, but as an essential element of the instructional process in physical education. Those findings had been supported in research results obtained by Lombardo and Cheffers (1983) who's employed the Cheffers Adaptation of Flanders Interaction Analysis System (CAFIAS). Also based on results of DPA Helsinki project, which used taxonomies developed by Flanders, Bales and Bellack separately and summarized was found the content to be a significant element of the instructional process related to the way of learning cognitive proprieties of verbal communication (Koskenniemi 1981). Reponen (1979) used the PEIAC/LH-75 (Heinilä 1977a) instrument in a normal setting (n=44x 20 min). Physical Education classes, were coded from videotaped material (Scott's Pi .79) to investigate direct-nondirect teaching behavior: It was established that (1) the order of indices revealed differences between two experienced teachers (n=24 classes taken from the studies of Heinilä 1977a) with regard to the rank order of behaviors, (2) the order of indices revealed differences between three female (3 x 12) and three male (3 x 9) teaching capability groups of P.E. students groups of student teachers (n=54), and (3) the order of indices distinquished between the two experienced teachers and student teachers. These results showed that the employed PEIAC/LH-75 system and measuring instrument could be used for the study of teacher -pupil interaction process in physical education classes (Reponen 1979, 111).

The results of this study seemed to ferity the construct validity and sensitivity of the developed PEIAC/LH-75 system and its observation instrument.

In present study, also the inverse character of reliability and validity was highlighted, which had already been pointed out by Flanders (1967, 1970) in his analysis concerning the training of observers and reliability problems (Heinilä 1980). Resolutions for implementation this kind of "coder validity" problem are multidimensional: firstly, the use of a multidimensional coding needs special training methods and qualified technical equipments. Secondly, differences in the 'validaties' of different recorders using the same system seems to be, in according to Medley (1982, 1841), in part a matter of aptitude and in part a matter of training practice and as he states: "nothing is really known about the attributes that make a person learn to code more easily, but they do exist". Thus, observation is a skill and its learning might be related to characteristics of a person, his knowledge, attitudes and expectations, as well as the use of this skill

in variable situation and for different purposes (Barrett 1983, Barrette 1996, Cloes et. al. 1995, Flanders 1967a, Heinilä 1980, Pitkänen et. al. 1979).

According to Locke (1977), "possession of reliable instruments for observation and knowing how best to use them, do not in themselves guarantee either sound research or fruitful results, but in the area of teaching they are essential first steps. And as we move to evaluative studies, we will have to confront the problem of multiple criterion measures and we will need product batteries which permit multivariate designs". This study has been an attempt to proceed in the direction recommended by Locke.

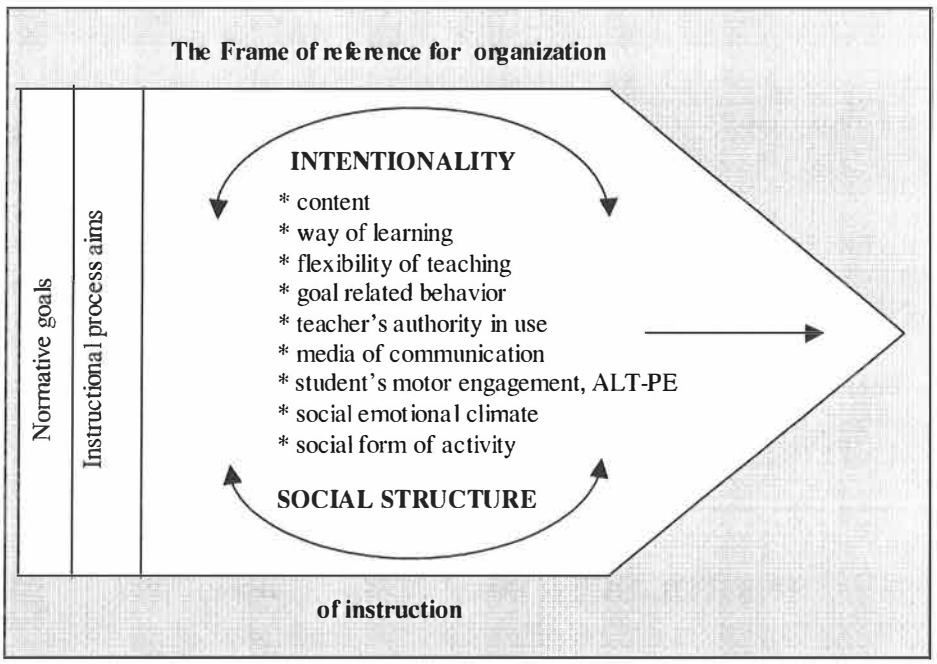
6.7 Activity forms in the paradigm of PEIAC/LH-75- system

On the basis of investigations for the development and validation of a System (PEIAC/LH-75) designed to expand the Flanders System of Interaction Analysis for Describing Teacher-student Interaction Process in Physical Education classes, also a paradigm of the instructional processes has emerged, a paradigm in which intentionality and social structure are considered the main elements of that process.

Intentionality is represented by the content and the way of learning (cognitive, affective and psychomotor proprities of verbal and nonverbal communication). Teachers' authority in use as well as flexibility of teaching are connected with intentionality and are seen as central aspects (Flanders 1965, 1970).

The other main element, the social structure of the process of instruction, in physical education, manifests itself as the teacher and student roles, which regulate the interaction process. These roles are reflected in division of labor and responsibility between the teacher and student, and in the grouping of students. The other characteristics of the interaction process in connection with the PEIAC/LH-75- System are the social emotional climate of gymnasium and student's cognitive, affective and psychomotor engagement called ALT-PE, Academic Learning Time in Physical Education.

The main elements of the goal-directed instructional Intentionality and Social Structure process are related to each other. These relationships can be seen illustrated in Figure 20 as process, in which the activity contents and forms of the instruction are represented.



(Heinilä 2002)

FIGURE 20 Activity forms in the paradigm of PEIAC/LH-75 (Heinilä 1977a)

SECTION II

THE APPLICATION OF INTERACTION ANALYSIS TO TEACHER TRAINING, IN PHYSICAL EDUCATION AND PROGRAM EVALUATION

7 INTRODUCTION

7.1 Background and need for the study

In Section I, Chapter I, it was stated that the central task of the university is the planning and realization of educational programs, the ultimate aim being the quantitative and qualitative development of education. The development of educational programs should be based on scientific research. The preceding chapters have reported the results of a study of physical education classroom interaction and the development of an observation instrument, which will permit a detailed description, and careful analysis of this interaction.

In this Section II, it will be reported how this research has resulted in a program of teacher education, which makes use of the observation instrument as a part of the training of future teachers of physical education.

It contains three successive phases which aim at the development and evaluation of teacher training program in physical education, more specifically a Flanders'-inspired didactic observation and microteaching course, based on experimentation and framework presented in Section I, (see Chapter I, p. 19 - 22 and Chapter 6, p. 32 - 39). The aim of the course was increased knowledge and mastery as well as cognitive understanding of characteristics of teacher-student interaction as defined by the author's adaptation of Flanders' Interaction Analysis System, PEIAC/LH-75, II (Heinilä 1977b, 1988, 1990, 1992a).

As already described in Section I, research on teacher education (ROTE) refers to research, which deals with any aspects of the process designed to influence what

teachers do in the execution of professional tasks. But, ROTE is also research on helping teachers acquire or improve teaching skills (see Locke 1983, 286).

Research in teacher education focuses on the pedagogical system, presage, context, process and product as immediate and long-term student growth. The most important question in program evaluation and assessment we must ask is, does the program have some conceptual integrity and coherence and does it is congruent with what we are doing in the program.

According to Dunkin (1987) "The teacher education context is especially interesting, because teaching is basis of the objectives guiding teacher education programmes, as well as process by which those objectives are attained and the main outcome by which the success of programme is judged (Dunkin 1987, 8)". (See Research review in section I, chapter 1 and 2). Therefore the basic assumption in planning of the program is that the concepts, models of teaching, methods and procedures used and validated can be applied to the teacher training program and that teaching skills based on models and experimentations can be taught (Borg et al. 1970, Everston & Green 1986, Flanders 1966, 1970, Gage 1978, Locke 1983, Rosenshine 1976).

Secondly, if intentionality of the instructional process were to be taken into account, the question arose how variables representing goal-oriented behavior are to be operationalized in the different phases of the instructional process. Thus, what is the way of learning the cognitive, affective and psychomotor properties of the program content?

In the early 1960s' a large number of teacher training packages, "minicourses" were developed from the concepts first presented by Dwight Allen and his colleagues at Stanford University. The best known of these are the minicourses developed in the Far West Laboratory of Education by Flanders (1966, 1970). They have also been used most often as a theoretical basis and methodological means in systematic teacher training programmes for physical education teacher education (Piéron & Cheffers 1988).

Several assumptions underlie all these minicourses. The most obvious assumption is that they can make teachers better and more effective (Flanders 1970, Gage 1978, Joyce et al. 1981). A second common assumption is that teaching skills are hierarchical and therefore there must be a sequence for training them. Coupled with this is the assumption that knowledge of the rational for a new behavior or the theory that supports the behavior is a factor in students learning of a specific new teaching skill. Thus, the more students know and understand the behavior to be learned, the better she or he will acquire it.

One important problem relates to the different types of minicourses in the acquisition of the same kind of teaching skill. Some of the applied extended, combined training programmes in physical education setting are extremely complex, requiring the student to spend 40-70 hours working through the course while others are relatively simple. Firstly, time is an important factor connected to all microteaching and minicourses. Therefore the educational institutions social settings where the course program is realized must be concerned with the relative time required by different teacher-training courses. Also the environmental facilities as materials, equipments, films and videotape

components, personal assistance needed for them must be considered. Time and cost effectiveness cannot be ignored on practical grounds, since they are both finite components. Further there is also the basic factor connected to the social structure of the instructional process: the group size in microteaching is a factor assumed to reflect the instructional process and acquisition of specific teaching skills. In physical education setting, as stated in the Framework (Heinilä 1977a, 1987, see Chapter 6, p. 170), the social form, division of the labor and responsibility and grouping of students is an essential aspect of the instructional process and therefore it cannot be ignored. It can also be asked whether a single gender group is more effective than a mixed group in microteaching setting by using student peers as pupils.

The content of the course needs also to be considered. Less complex of subtle teaching skills might be learned in simpler training methods, whereas more complex, theory based teaching skills and strategy/to be learned needs specific approach like cognitive oriented models, or/and models of experimental learning (See Flanders 1970, Hanke 1980b, 1987, Hanke & Treutlein 1983, Hytönen & Komulainen 1971, Keilty 1975, Kirk 1986, 1993, Rogers 1967, 1980).

Furthermore, if intentionality and social setting of the instructional process is taken into account, the question arises, how the program can be integrated to the contextual frame, and the curriculum of the faculty. Based on assumption that teacher education is a life long process and that students entering these programs come with a specific background, the timing and coordination of the program to curriculum cannot be ignored. Teacher education programs at the university level must emphasize the process where a student becomes day after day a better teacher by making the acquisition and integration of knowledge, abilities and attitudes related to educational goals of the curriculum (e.g. Feingold & Barrette 1988, Heinilä 1988, Hupé 1995, Lawson 1988, Telama 1968, 1970, Telama et al. 1988, Telama & Vuolle 1976).

In program evaluation, the impact of the program for different students and course groups is of central concern and needs to be considered. The evaluation would not be complete without the self-evaluation by the teacher educator of the education strategies experienced by students with respect to the pertinence of the suggested activity framework, the content and form of the teacher training program and with respect to the pertinence of the real contribution of the teacher educator regarding the needs of students in connection with the different phases of the instructional process.

Finally, we are faced with an important and neglected problem, recognized also in research reviewers, related to the contextual factors in acquisition of different kind of teaching skills and strategy. As Locke (1983) the states:

> "As a body of knowledge and domain for inquiry in physical education remains uneven, unpopular, and unread"...and that, ..."our near total failure to examine the social and psychological context of teacher education from the perspective of the participants is the main impediment to its improvement" (285).

Whether it is connected with a practical issue, such as time sequence, social structure, group size, or assistance needed, or a theoretical issue connected with the content, the main elements of the program and the nature of learning (the cognitive, affective and psychomotor properties of the content), there is a need for research on the internal and external validity of training programs.

7.2 Microteaching and didactic observation in teacher education curricula

As stated in Section I, Chapter I, in January 1974, the Department of Physical Education of the Faculty of Health and Physical Education at the University of Jyväskylä introduced, on an experimental basis, a new type of practice teaching in the form of a course on microteaching. It formed part of the degree requirements and was given during the last term of the third year as an obligatory course (45 hours, lectures 15 h, practice 30 h (Telama 1975). The experiment was started as a result of the positive reports on the use of microteaching and interaction analysing systems as a tool of teacher education (cf. Flanders 1970, Heinilä 1971, 1974). It was considered to have a potentially beneficial effect on the attainment of the objectives of teacher education in physical education as well as on bridging the gap between the theory and practice of teaching. It was for the implementation of this course that the interaction model and observation instrument, PEIAC/LH-75, was constructed. The measuring device had been pilot-tested at the beginning of the course and its use, in modified form, PEIAC/LH-75 II proved feasible (Heinilä 1977b).

Although observation has been consistently seen as an important skill for teachers (see Barrett 1983, Borg et al. 1970, Flanders 1970, Komulainen 1974b, Koskenniemi & Hälinen 1970, Wagner 1971) and especially for coaches and teachers of physical education (see Barrett 1979a, Pitkänen et al. 1979, Telama et al. 1980), it had not had a recognized role in the P.E. teacher education programs nor on research on teaching in the early 1970's. However, at the same time of starting the course of microteaching, a course of didactic observation was introduced to the curriculum of the Department of Physical Education (30 hours, 15 h lectures, 15 h demonstrations), given during the first term of the third year and with the aim to develop student's observation skills and make them familiar with teaching through a systematic analysis and experimentation. This course also teaches evaluation since it deals with the evaluation of the interaction process (Report of the 1973 Commission on Teacher Education, Report of the Commission of Education 1975:75, Telama 1975, 1978, 1979, Telama et al. 1980 and Heinilä 1977b, 1988). These courses were developed and taught by the author, Faculty member until the year 1991.

Teacher training process in Jyväskylä University in the early 1970's resembled that in other countries. It was training divided into two levels: a first level, called pre-service, which included all of the operations used to prepare students for entry into a teaching career (higher education with "limited

responsibility" teaching) and second - level called in-service training - which includes all of the operations used to continue the education of teachers after they are certified for teaching in a "Normal School".

Also the observation-based subjective rating of pre-service student's achievement proved to be rather unsatisfactory because no research model representing the instructional process as a whole was available. This lack was reflected in difficulties to evaluate the process of teacher training and to integrate the theory and practice in teacher education as prescribed in the normative curriculum of the Faculty. (Heinilä 1977b, Telama 1975, Telama & Vuolle 1976).

When the earlier forms of practice teaching, so-called order-calling exercises, were given up as not being congruent with the principles of the new type of P.E. teacher education, there was a decrease in the amount of practice teaching. The student teachers felt that this was a disadvantage, leading to a feeling of uncertainty when they started their one-year practice teaching in the in-service training at the "normal school". The need for new opportunities for practicing was clearly indicated.

The training of physical education students in Finland changed, however, in the study reform 1978 with the following characteristics underlined: (1) problem centered orientation, (2) scientific orientation, (3) multidisciplinary orientation, (4) vocational orientation. The main differences between the earlier training of physical education teachers and the new system was that the new educational program was more goal oriented and the structure of study program was different from the earlier one (the new degree comprised 160 study-week for students; one study-week is 40 hours) and the work spent on to the studies by the student was greater, which means a longer total study time – an average of 5-5½ years. The link between theory and practice was more emphasized than before, and most evident in connection with practice teaching. (Asetus liikuntatieteellisistä tutkinnoista no. 299 (Requirements of science of P.E. examinations) Ministry of Education (21.4.1978), Heinilä 1988, Telama et al. 1980).

In connection with study degree Program reform at the Department of Physical Education (1978) also the courses of Didactic Observation and Microteaching reorganized in the following way: *firstly, the two courses were combined in degree requirements, and formed two study-weeks (80 h) united and assessed together (1-3 p); secondly, the time reservation for the demonstrations and practice was however diminished 13 hours (course of didactic observation 27 h; 15 h lectures, 12 h demonstrations; course of microteaching 35 h (15 h lectures, 20 h demonstrations); thirdly, the timing of these courses was changed to the earlier stage in the study program, beginning with the course of didactic observation in the least term of the second year and continuing with the course of microteaching in the first terms of the third year.* (The study guide of the Department of Physical Education 1979-1980, 403.01.EA, 72). However, it can be noted, based on information given in the documents of the Faculty that the aims, contents and forms of the revised course package did not change and they were congruent with the general principles presented in connection with the study reform. Therefore, also the revised course package was expected to be at least as effective as the earlier course program in the future P.E. teachers' training curriculum. Also the course

organization was changed so that they could select own group membership and practice microteaching in mixed gender groups. And moreover, it might be noted that the important contextual factor, *the Student Intake Procedure to the Department of Physical Education was changed;* e.g. the test of students entry teaching skills, teaching episode tests weight in total scores, was diminished from 25 percent to 15 percent, and in 1980 to 11.5 percent (Appendix 9.1, Heinilä 1988), the student population, their background variables were different from before the study degree program in general. Thus, a "reflective-oriented practice" course-package was conducted in seventeen years from 1974 to 1991 in two different kinds of contextual settings before and after the study reform and with the same aims and contents. The main point of interest was how the contextual variation affected congruence between objectives and the degree of their realization. To answer this question, an evaluation of the curriculum program and its realization in a longitudinal, multidimensional design was needed.

7.3 Evaluation of curricula

The scientific basis of teacher training consists of knowledge of regular, no change relationships, in the realm of events with which the practice is conducted (Gage 1977, 15). The *"curriculum"* refers here to a plan of all measures undertaken in order to attain set objectives. *Evaluation* is considered to be activity the purpose of which is to obtain information for making decisions between different alternatives (Gage & Berliner 1979, Heinilä 1977b, Stake 1967, 19, Stufflebeam 1968, Telama 1978, 1979, Worthen & Sanders 1987). Analysing alternative approaches and practical guidelines of educational evaluation Worthen and Sanders (1987, 130) noted that Stake's (1967) early analysis of evaluation process had a major impact on evaluation thinking and laid a simple but powerful conceptual foundation for later developments in evaluation theory. Stake's (1967, 529) countenance model illustrated in Figures 21 and 22 is used as the frame of reference in this curriculum evaluation study (Heinilä 1977b).

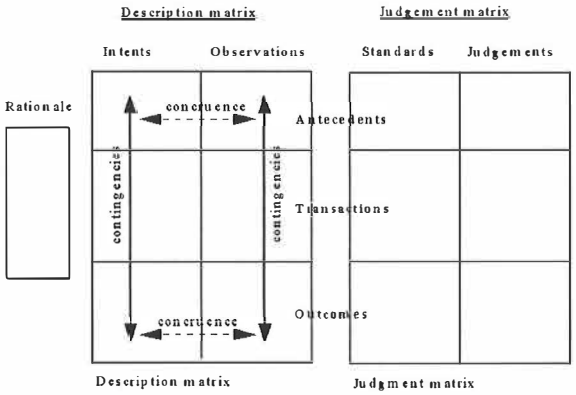


FIGURE 21 A layout of statements and data to be collected by the evaluator of an educational program (Stake 1967, 529)

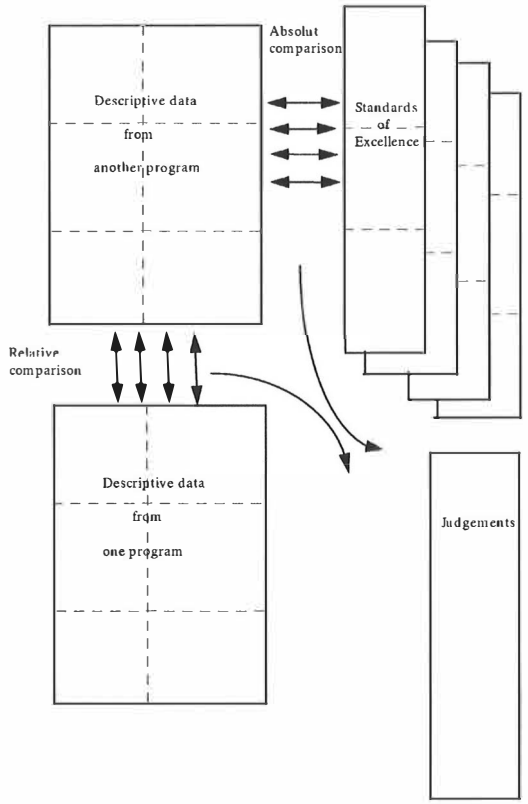


FIGURE 22 A representation of the process of judging the merit of an educational program (Stake 1967, 230)

The main activities of descriptive -judgemental evaluation are (1) the study of the contingencies of antecedents (inputs, resources, existing conditions), transactions (activities, processes) and outcomes, and (2) the study of the congruence between the level of objectives and the level of observations. Congruence indicates to what extent the plan is being carried out (Stake 1967). During the early 1970's Stake expanded his earlier (1967) countenance model more obviously into the real or naturalistic and participant-oriented evaluation. In "responsive evaluation", Stake (1978) stressed the importance of the evaluator being responsive to realities in program and reactions, concerns and issues of participants rather than being "preordinate" with evaluation plans, relying on formal plans and objectives of program. Stake defined responsive evaluation as follows:

> "An educational evaluation is responsive evaluation" if it orients more directly to program activities than program intents; responds to audience's requirements for information; and if different value perspectives present are referred to in reporting success and failure of the program (Stake 1975a, p. 14, Worther & Sanders 1987, 134).

8 AIMS OF THE EVALUATION STUDY

The purpose of the evaluation study was to create, by means of literature survey and empirical research, a basis for the development of new forms and contents of practice teaching so that they satisfy the demands of changing physical culture on teacher education and to find a suitable place in the overall educational program for them.

In the first phase of this curriculum evaluation, educational intentions at the curriculum level and their degree of realization at the observation level will be the object of research. In the third phase, the model of responsive evaluation will be applied as the frame of reference.

Section II will present some descriptive-analytical, predictive and explanatory information with the aim to answer the following questions connected to the sequential phases of program development, control, implementation and long-term evaluation:

(1) What is the rationale behind the didactic observation and microteaching course and how is the course implemented in practice? How is it related to the observation method developed for analysing the interaction process in physical education teacher training? (Content and context variables)

(2) How does the microteaching programme work in practice? What is the degree of congruence between intented antecedents and what actually occurs in two microteaching settings which differed with regard to (1) modelling sequencing of teaching, (2) timing, (3) size of group, and (5) number of reteaching? (Change in student's teaching behavior in criterion task described in expectations of performance of comparable programs) and in student's experiences? (Content and process variables)

(3) What is the level of applicability, accountability and construct validity of the basic elements of the program: modified PEIAC/LH-75, II, observation system as feedback instrument with teaching skills, operationalized as indirect teaching models, and the reliabilities and validities of the other elements of the rating scales used as means of intervention and student's course rating.

(4) How does the program serve students with different presage variables (students entry characteristics evaluated in student's intake didactic

observation course and before the microteaching course)? (Context-presage-program-process and output variables).

(5) How does the didactic observation and microteaching course serve the goals set for the practice-training period and how is it integrated with the normative curriculum of the Faculty in contextual variation? What is the external validity of the program? (Replicated context, presage-process output investigations).

(6) How does the program serve male and female students before and after the study degree program reform (1978)? (Student's program evaluation)(In program predictive validation the interest is more on the criterion, thus in students' achievements than in predictor variables and in students ratings of the program, based on their own experiences).

9 THE FRAME OF REFERENCE

The program development and research work involved the components and sequential phases illustrated in Figure 23. The study contains three successive phases: the two first are concerned to meta-level problems, on concepts, methods, procedures and validation of the basic content elements of the program, whereas the final phases are concerned with problems at the substantive level in contextual variation.

In Phase I, two versions of a microteaching course are compared in order to assess the effectiveness of their components. For the purposes of this comparison, microteaching is described and its components are analyzed, particularly those on which this study focused. In the empirical part of the study, a short description is given of the teaching program, design, research tasks and methods of measurement and analysis. Also results of inquiry into students' ratings of the two-microteaching course are described, analyzed and compared. The results of the explorative study are then presented and discussed.

In Phase II, the construct validity and sensitivity of the basic content elements of the revised program is analyzed and estimated in pilot studies by using a multivariable approach: (A) the modified PEIAC/LH-75 II-system used as a feedback instrument in microteaching by using the teaching models (1-6), operationalized teaching skills (Heinilä 1977b, 1990); (B): the PEIAC/LH-75 II-system as a research instrument and means to observe the sequential processes in microteaching setting: then the evaluation-revision cycle is repeated to assess the stability of the factorial construct and factor structures, reliabilities and validities of the scales and questionnaires; (C) a rating scale for determining students' entry teaching skills, (D) a questionnaire for assessing students' attitudes, "ideal" P.E. teacher expectations, (E) a questionnaire concerning student program evaluation.

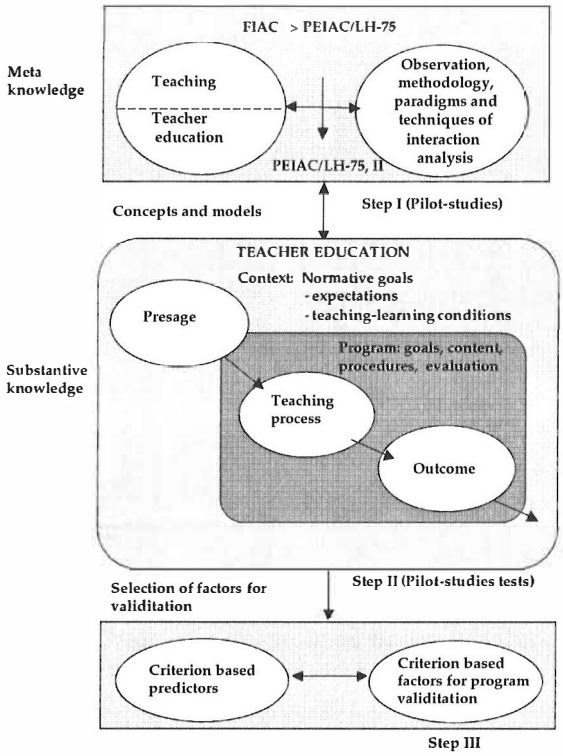


FIGURE 23 The frame of reference: main components, in relation to each other and research strategy (Heinilä 1992a)

Phase III of the study is a long-term longitudinal empirical ex-post facto inquiry for estimating the internal and external predictive validity of the program. A multiple research approach with constant research problems, repeated investigations and replicated designs is used. The study is conducted in the frame of a contextual setting by variation of curriculum and student selection procedures in a long period of time 1974-1988. Predictor variables consisted of students' entry characteristics such as gender, students' intake scores and course program intervention variables, students' attitudes and entry teaching behavior. The criterion variables consisted of students' achievement variables, such as measured teaching behavior and final grades on the course theory, practice and total scores. Secondly, for the purpose of program validation, also the results of the students' program evaluation in the contextual frame are analysed and compared in order to assess the effects of contextual variation on student ratings in program "implementation" dimensions. It is assumed that the research knowledge provided is useful for the implementation of teaching-learning conditions to attain objectives of the program as prescribed in the normative curriculum of the faculty. In the final section, some recommendations are given for the course of didactic observation and microteaching in the P.E. teacher education program, as well as for related follow-up and research activities.

10 REVIEW OF RESEARCH

10.1 Forms of practice teaching

The development of research on teaching and the methods of interaction analysis have been discussed in Section I in Chapter 5 of this dissertation with special attention to Gage's (1972) model of the field of research on teaching. The adapted version of this model was introduced in Section I, Chapter 1 (see Figure 2, p. 22). It illustrated the place of the present discussion within the research area and described the starting point for the study of teaching programs (Heinilä 1977b). In this section, the concepts and purposes of microteaching and minicourses and some research results obtained in connection with the application of them to teacher training in physical education are reviewed.

Microteaching

When microteaching was first planned at Stanford University in 1963, the aim was to develop more effective forms of practice teaching. The following criteria for the organization of initial training for preintern teachers were set (Allen & Clark 1967):

First: A real teaching situation was needed so that candidates could be actively engaged in practicing and refining teaching skills and experimenting with their own and their supervisors' ideas.

Second: The teaching situation must keep the risk low both for the teacher and the students.

Third: The pre-service teaching context should take into account some well-established facts within learning theory. For example, numerous distributed practice sessions; immediate supervisory feedback; immediate opportunity to rectify errors and weaknesses; low anxiety, etc.

Fourth: The pre-service context should provide a setting in which the trainee can have experience with a wide range of student abilities and age

levels and develop competence with a broad spectrum of teaching skills.

Fifth: Economy in terms of time and resources should be maximized.

Microteaching was conceived to meet these criteria.

The spread of microteaching into colleges of education was very rapid. In 1972, 50% of the colleges of education in the U.S. used various adaptations of microteaching.

Allen and Eve (1968) define microteaching as "a system of controlled practice that makes it possible to focus on specific teaching behaviors" (p. 181). The term "system" here, as well as in discussions of systematic observation, refers to the rigorous plan of choosing and controlling the components of the system beforehand for a certain specific purpose.

In microteaching, the teaching situation is usually scaled down in terms of time and number of students. The "session" lasts four to twenty minutes and the number of students varies from three to ten. Microteaching can be used for a number of purposes. Some of the variables, which can be adjusted, include lesson length, number of students, type of students, number of "reteachings", the amount and kind of supervision, and the use of videotape (Allen & Clark 1967).

Microteaching's component-skill approach is used primarily to give the trainee a clear idea of the skill to be learned. The trainee has to know what he should do before he tries to do it. Instruction in a particular skill can be given by oral instructions, written directions, demonstrations or combinations of these. In the usual Stanford microteaching sessions, the procedure is to teach 5 minutes, critique 10 minutes, replay 15 minutes and reteach 5 minutes (Allen & Ryan 1969).

Minicourses combine some of the features of microteaching such as practicing model learning and the use of feedback derived from the observation of the video tapings. Furthermore, some characteristics of programmed instruction are evident, for instance, in independent learning.

Research and teaching training have paid increasing attention to the component skills of teaching. Borg and his colleagues (1970) at the Far West Laboratory developed some of 20 self-instructional minicourses designed to train teachers in the use of specific classroom skill, such as questioning, organizing independent learning etc. Each minicourse was produced as a result of rigorous development and research work which involved the following components and sequential phases; (1) the stating of specific objectives for the product; (2) the use of available research knowledge as source of concepts and materials; (3) the carrying out of field testing programmed to evaluate the product effectiveness in a setting where it was eventually to be used; (4) the use of results of this evaluation programmed to improve the product; (5) the evaluation programmed to improve the product; (6) the evaluation-revision cycle being repeated until the product met its prescripted objectives. Each of these minicourses made use of a systems approach with steps occurring in the following sequential fashion: (1) precise specification of the behavior which is the objective of learning experience; (2) carefully planned training procedures

aimed explicitly at those objectives; (3) measurement of results of training in terms of behavioral objectives; (4) feedback of the observed results; (5) re-entry into the training procedure; (6) measurement again of results (Borg et al. 1970).

According to Flanders (1987, 26) Borg's model for minicourses emphasizes, firstly, reported feedback from the cycle of microteaching and the need for modelling teaching, and secondly, careful field testing during the development of instructional material for teachers.

In order to make minicourses as effective as possible it is of particular interest to study the effects of its various component factors. The problems are partly identical to those encountered in the development of adaptations of microteaching, the best known of which is the minicourse developed in the Far West Laboratory of Education by Flanders (Flanders 1966, 1970). It is a teaching package consisting of sound films and printed materials, which present the model and instructions.

In late 1970's, about a dozen of minicourses were developed, tested and marketed. The testing consisted of experiments in which teachers were observed systematically to determine the pre-existing level of teaching skills at which the minicourse was aimed. The catalogue developed by Gage and his associates at Stanford contained more than 800 (650 titles) products (Gage 1978). A teacher-training product was defined as material intented to equip teachers which skills or knowledge of "how" to do certain things rather than knowledge, that certain things are true. The Stanford catalogue described the hundreds of products in terms of the following nine dimensions: (1) the product's subject matter specifity; (2) the target audience; (3) the grade-level specifity; (4) the so-called target outcome; (5) the target outcomes for students; (6) the training situation; (7) the time and number of persons required to administer training with the product; (8) the kind of practice prioritized; (9) the phase of teaching in which the acquired skills would be used. Ten years later Joyce and Weil (1980) presented, in a book of "Models of Teaching", 25 models divided into four families. There were models that emphasize social interaction, information processing, personal development, and behavior modification cybernetics. Joyce et al. (1981) believe that teachers need to know one or two models in each family to build the repertoire required for flexible teaching. Several assumptions underlie all these materials. The most obvious assumption is that they can make teachers better and teaching more effective (Gage 1978, Joyce et al. 1981).

10.2 Contents of practice teaching

The skills chosen as targets of practice in the new type of practice teaching programs in the 1970' varied with regard to their degree of specificity and concreteness, cognitive level, the theory on which the choice has been based, etc., in accordance with the set objectives, forms of teaching and resources (Gage 1979).

Criteria in the selection of patterns are, for instance, their relationships with student learning: knowledge, skills and attitudes. Which of them we choose to stress in P.E. teaching is a question that is related to our conception of physical education in general. It should be noted that effectiveness thinking is not the same as process-centered thinking. Often expressiveness is a condition for attaining instrumental objectives (see, e.g., Bookhout 1967) - at least in physical education - in which social form, nonverbal communication, peoples' engagement motor activity (ALT-PE) and the affective element are also emphasized (Cheffers 1977, Hanke 1976, 1980b, Heinilä 1977b).

In 1965, isolated technical teaching skills were practiced in the Stanford Laboratory of Microteaching, including initiation, presentation (communication), consolidation (of the lesson), monitoring and evaluation (Allen, Fortune & Cooper 1967, Brusling 1974). They are similar to the basic characteristics of the phase of the social interaction process (see e.g. Bales & Strodbeck 1967). Allen and Ryan (1969) also give a list of general skills amenable to practice whose application to the teaching of different subjects and different levels of pupils is possible: (1) stimulus variation, (2) set induction, (3) closure, (4) silence and nonverbal cues, (5) reinforcement of student participation, (6) fluency of asking questions, (7) probing questions, (8) higher-order questions, (9) divergent questions, (10) recognizing attending behavior, (11) illustrating and using examples, (12) lecturing, (13) planned repetition, (14) completeness of communication.

According to Dunkin (1987), the definition of technical skills of teaching had broadened in more recent years into diagnostic, analytic hypothesizing skills rather than more strictly observable skills. The emphasis had tended to move more towards teachers' powers of conceptualising, hypothesizing and synthesizing than to the acquisition and performance of discrete teaching behaviours through technical skills of teaching.

In connection with the use of interaction analysis, these component-skills refer to the sequence of teacher-pupil interaction and are called "patterns (or models) of teaching". A pattern is a short chain of events that can be identified, occurs frequently enough to be of interest, and can be given a label (or name) since this often facilitates thinking" (Flanders 1970, 4).

There are, however, still problems regarding the validity of many specific teaching skills included in teacher training programs and suggested criteria for making judgement about student teachers' learning.

As Dunkin (1976) argued "two criteria should be applied in judging the validity of technical skills of teaching. The first is the extent to which the specific aspects of teaching behavior is distinct from other aspects of teaching. Observers should be able to agree on what constitutes the nature of the skill and should be able to identify it when it occurs. The second criterion is the extent to which the skill can be shown to enhance students' learning". Dunkin (1987, 705-706) noted further that, "such assumptions about the absolute validity of some teaching skills are not recognized by supporters and users of the approach, where-as defenders of the approach argue that technical skills of teaching are best incorporated into teacher education programmes not nearly as recipes for action but as a part of behavioural repertoires which heighten teacher's

capacities to select and implement teaching strategies they might choose on the basis of their theoretical training and creativity dispositions".

The later conception is close to what Flanders (1987, 22-27) argued in his critical comments concerning process-product research connected with the Human Interaction Analysis Model used as frame in teacher education - with the objective implementation of teacher's behavior, flexibility in authority in use. However, Flanders (1987b, 460-465) also observed that the problem of validity was still of central importance in research on teacher education.

10.3 Some research results

10.3.1 Some research results on teacher education (ROTE)

Flanders (1970) reports 18 research projects, which investigated at different levels of education the effectiveness of using interaction analysis as a means to facilitate learning. A general objective of such programs was an awareness of teaching behavior and the development of flexible teaching behavior. Research findings summarized in Flanders (1970) gave rise to some generalizations:

1. An individual becomes more responsive to pupil ideas, the amount of open and higher-order questions increases, statement of reasons increases in connection with praise and criticism.
2. Teaching behavior becomes more flexible or variable and more guided by situational factors.
3. The attitudes of student teachers toward the new type of practice teaching become more positive.

Flanders stated that "interaction analysis can help to develop value systems about teaching which we call convictions, by contributing information which is primarily objective" (Flanders 1970, 19).

In research connected to microteaching in the early 1970's, it was recognized that variables, such as content of teaching, skills, length of microlessons, size of classes, the teach-critique-reteach cycle patterns of supervision and the use of models had been the subjects of extensive experimentation and research in the search for optimal procedures, although many of those tested had already yielded significant results (Borg et al. 1970). Microteaching is, as stated in introduction, based on a long-established learning theory which today underlies programmed learning and computer assisted instruction. It is also assumed that learning is more effective if complex skills are divided into their components and learned step by step before it is undertaken as a whole. Other contents of learning theory such as feedback, reinforcement or extinction were also adopted in the microteaching procedure. Observational learning through a "model" was found to be another example of well established educational theory and practice (Bandura 1969, Borg et al. 1970, Dunkin 1987, Flanders 1970, 1987, Joyce et al. 1981, Wagner 1971). Discrimination training had been found to be the most important component in microteaching (Borg et al. 1970; Wagner 1971). In an experimental study,

Wagner (1971) found that training in discrimination classes of pupil centered teacher behavior produced more such behavior than repeated practice in microteaching setting. However, there was no comparison with a group which had both microteaching and discrimination training. The research studies referred and summarized by Rosenshine (1976) represented methodological and conceptual expansion and focused e.g. on student behavior in terms of concepts like Academic Learning Time (ALT), time on task and students' motor engagement time variables, which had been found to be of central importance in connection with physical education. (e.g. Carreiro da Costa & Piéron 1990, Dodds & Rife 1983, Heinilä 1971, 1977a, 1977b, Piéron 1982a, 1996, Piéron & Piron 1981)

10.3.2 Some research results on teacher education for physical education (ROTE-PE)

The contents and forms of practice teaching in physical education was studied relatively little in the early 1970s'. The need to develop new types of practice teaching along the performance-based teacher education lines was recognized (e.g. Cheffers 1977, Feingold 1972, Finske 1967, Hanke 1976, Heinilä 1976, Jawett & Müllan 1972, Lundgren 1972, Siedentop 1972).

It has been stated in Section I that Flanders' Interaction Analysis system FIAC has been used most often as the feedback instrument in connection with systematic physical education teacher training (e.g. Akkanen 1979, Barrett 1971, Dougherty 1983, Garrett 1973, Hanke 1976, 1980b, Harrington 1974, Heinilä 1977b, Mancini & Cheffers 1983, Mancuso 1972, Melograno 1971, Piéron & Cheffers 1988, Splinter et al. 1979, Steward 1977, Reponen 1979).

Exploratory studies of teaching behaviors in physical education (e.g. Cheffers & Mancini 1978, Dougherty 1970, Hanke 1976, Heinilä 1971, 1974, Nygaard 1971), which used observation instruments derived from Flanders' FIAC system, found that the behavior of teachers and student teachers in physical education was direct (teacher-centered). Typical of P.E. teacher's speech behavior was also the lack of variation in terms of the features of social interaction and the dominance of teacher talk (e.g. Cheffers 1973, Heinilä 1971, 1974, Reponen 1979).

In the area of teacher training two main research designs were used: (1) experimental studies and (2) descriptive studies:

(1) In physical education teacher training, the most common objectives of the experimental studies in 1980's had been to investigate effectiveness of different teaching strategies, learning of teaching skills and teaching styles, effectiveness of the training programmes, effectiveness of different feedback systems used (see Hanke 1980b, Locke 1983, Piéron 1982a, 1996, Piéron & Cheffers 1988). Many studies focused on student motor engagement time (MET) or academic learning time (ALT-PE) in the microteaching setting as initiated by findings summarized in process-product studies by Piéron (1982a, 1996), Piéron & Haan (1980), Piéron & Piron (1981). Borys (1986b) developed a training procedure to increase pupil motor engagement time (MET). The findings of the descriptive

inquiry showed that during reteach lessons, treatment teachers as a group showed greater gains in pupil MET occurrence (from 43.4% to 50.8%) than the control teachers (from 40.2% to 42.5%). Carreiro da Costa & Piéron (1990b) pointed out the role of the teacher as facilator of learning, the importance of 1) time-on task at a high level of success, 2) specific and correct feedback and 3) instruction provided by teacher. It was found that teaching learning variables were related to student success in the experimental unit used. Also studies had been designed for working out intervention procedures included in teacher training course packages. It had also been found that several interventions were more efficient than the use of only one type of intervention (Siedentop 1981).

Hanke (1980b, 1986) summarized the results of European and North American investigations on the structure and effectiveness of different types of physical education teacher training studies up to 1976. The trend analysis was based on 400 explicitly empirical studies. He concluded that the trend of the development of the research designs was obvious: more and more different aspects of behavior modification were included as well as more different types of modelling were applied. In general, the audio-visual presentation of teaching models showed to be more effective in training teaching skills than verbal or written material. Hanke (1980b) summarized in his dissertation the results of 19 selected studies where the modifications of programs concerning, modelling, cueing and personal AV-feedback variation were studied as follows:

1) interaction analysis feedback was much more effective than unstructured generalizing remarks;
2) if interaction analysis feedback was combined with additional feedback-sources (AV, peer, supervisior, computerized interaction analysis such as "ratios" and "frequencies"), this combination showed stronger effects than the use of one feedback-method by itself;
3) the analysis of one's own behavior was more effective than the analysis of behavior of others:
4) through the use of interaction analysis systems apparently the subjects are cognitively influenced concerning the desired ability to teacher behavior (97)

In most of these studies (11) the Flanders' FIAC-system in modified and applied forms was used as a feedback instrument.

Based on these findings, Hanke (1976, 1980b) focused on cognitive discrimination training and behavioral change. He developed a micro-teaching program by using as a feedback instrument the 22-category "Heidelberg Interaction Analyse fur Sportunterricht (HIAS)" (Hanke 1976) which is an adaptation of the Flanders FIAC (1970). The experimental study, program evaluation, was focused on the comparison of the effectiveness of feedback of different treatment groups. Differences between eight groups were estimated and compared by using discriminant analyses technique, e.g. male and female groups were found to be significantly different in treatment gains.

However, it had been recognized that studies using the design experimental vs. control group had provided mixed results concerning the impact of these programmes (Piéron & Cheffers 1988). One can argue in agreement with Piéron (1992, 26) that in experimental studies differences in

coding procedures, selection of sample, and length of instruction are so large that comparison of these studies is meaningless.

(2) Research methodology by using descriptive research designs in connection with teacher training had been used more frequently in 1980's in studies connected to acquisition of teaching skills. The modification of the teaching behavior by using multiple baseline techniques had been found to be successful to investigate e.g. the causal relationship between intervention and change in specific teaching behavior or knowledge and to study the dynamics of behavior modification (Barrette 1977, Cheffers 1990, Currence 1977, Dougherty 1983, Heinilä 1977b, 1987, 1988, Locke 1986, Piéron 1996, Rogers 1980, Siedentop 1986). E.g. Lombardo and Cheffers (1983) employed a multiple observation approach and a modified case study design with the purpose to observe and describe the teaching behavior and interaction patterns of physical education student teachers "longitudinally" in microteaching settings as: one time before and two times in the course. This design was also used by Borg et al. (1970). The purpose of descriptive studies was to investigate how the program works in practice, how the students with different presage variables or how different groups passes through the programs. The impact of the programme for different groups in contextual variation is also important aspect of teacher education but it was found to be neglected in 1970's investigations as Gage & Berliner (1979) and in connection of physical education Locke (1983) states. The descriptive and analytic studies dealt with an increasing frequency also with beliefs and concerns of students candidates, their professional and or "ideal teacher" expectations (see Carreiro da Costa et al. 1995, Flanders 1987, Heinilä 1988, 1992b, Hendry 1969, Hytönen 1973, Hytönen & Komulainen 1971, Laakso 1975, Placek & Dodds 1988, Rogers 1967, 1980) and students' own experiences of the training program (Doolitte et al. 1993, Ebbs 1975, Heinilä 1977b, Rudy 1974, Schempp 1985, Telama et al. 1988). Placek and Dodds (1988) investigated teachers' beliefs and employed the critical incident method about success and non-success and as subjects 195 preservice student teachers (134 women and 61 men). 431 success features were identified. The categories used in comparison were: motivation, planning, feelings and methods. Success (16%) and enjoyment (15.3%) were found to be the predominant sub-categories in the student centered area.

Whitehead (1980) compared P.E. students' (n=1163) and other subject students' (n=554) personal characteristics in ten colleges of education in England and Wales and the predictive values of procedures used in selection procedures and of students' success in examinations. It was found (1) that P.E. students tended to obtain higher ratings at personality interviews and (2) higher rates in main subjects' final examinations and for teaching ability as students in the other subjects. P.E. students' success in examinations was revealed to be related to differing measures of extraversion, stability and tough-mindedness. A comparison in the time sequence of students and teachers professional socialization process had also received more attention in 1980s (Lawson 1988, Piéron 1996, Schempp 1985, Telama et al. 1988). Also the training in didactic observation had been investigated with increasing frequency, with

the purpose to verify whether that kind of teacher training contributes to modify teacher behavior or teacher-student interaction (Barrett 1983, Borg, et al. 1970, Cloes et al. 1995, Flanders 1970, Heinilä 1988, 1992a, 1992b, Piéron 1996, Pitkänen et al. 1979).

Many areas had, however, been neglected as Locke (1983) concluded in a summary review of research reports of P.E. published since 1980 in USA (Locke 1983, Locke & Dodds 1981). Little or no ROTE-PE had attended to following important topics:

(1) change and development in teacher education programs; (2) induction and socialization into the teaching role; (3) inservice teacher education; (4) career patterns and professional development; (5) recruitment and selection of trainees; (6) placement of graduates from training programs; (7) teacher educators; (8) the experience and perspectives of participants in teacher training programs; (9) control of teacher education in physical education; (10) the ROTE-PE enterprise itself conducted (292).

Moreover, in term of methodology selected for research and manipulation of knowledge base, ROTE-PE had been found to be involved little or not in the sample of studies analysed by Locke (1983, 293).

However, in late the 1980s, based on trend analyses connected to research in sport pedagogy, a clear evolution in ROTE-PE was visible in the areas research methodology and instrumentation (Bain 1990a, Carreiro da Costa 1993, Locke 1984, 1986, 1989, Piéron 1996, Silverman 1991, Silverman & Skonie 1997). Methodological problems and knowledge base of ROTE-PE was involved more frequently in research connected to teacher/coach preparation. Piéron (1996, 65) identified e.g., that of data-based selected studies, based on the sample of studies during 1980 – 1990, 30% represented teacher preparation issues but that only 15% of them were involved with curriculum problems. The main issue, acquisition of teaching skills, represented 50%. However, it was no more unique in this area: knowledge base of teaching and teacher preparation, values, attitudes, students socialization to teachers role longitudinally as a life long process were areas involved in ROTE-PE more as before. Based on analyses of 243 research documents, Piéron (1991, 67) concluded that observation was still in the 1990's research a classical means to gather data, and that also data collected by techniques related to teacher-pupil thinking had increased in noticeable proportion.

Also Silverman (1991) underlined in his trend analysis the importance of methodological changes to broaden the scope: "contextualizing, the use of ethnographical interpretative approaches techniques is needed to extent the perspective and to help to understanding physical education and teacher education from the perspective of teachers and students" and as he states "some questions about effectiveness and attitudes can only be investigated from this perspective" (Silverman 1991, 235-236). Also Feingold & Barette (1988) and Barrette (1996, 144) underscored this point in stating:

> "We need strategies which combat curricular fragmentation and faculty dissociation and which promote convergent strategies of planned integration connected to the real work of teachers in long overture".

This evolution issue was applied in the present also long-term study (Heinilä 1974, 1977b, 1987, 1992b).

10.4 Summary

The concept of specific teaching skills was implemented for the first time in teacher education in the microteaching program at Stanford University in the early 1960's. Microteaching is based on long-established learning theory, which underlines programmed learning and computer assistented instruction. Technical skills of teaching are specific aspects of teaching behavior that are considered to be particularly effective in facilitating desired learning of students. Most of studies focusing on chancing specific teaching behaviors have showed that the behaviors can be changed in teacher training courses. In connection with microteaching, Flanders' System has been applied most frequently and has been modified to a significant extent by varying coverage, method of data collection and coding procedures as well as the conceptual posture used. It is assumed that learning is more effective if a complex skill is divided into its components and learned step by step. Observation of one's own teaching behavior is more effective than observation of behavior of others. The concept of learning theory such as feedback, reinforcement or extinction were also adopted in microteaching and in the minicourses. Observational learning through a teaching model is one example connected e.g. to Flanders' (1970) well established educational theory and practice. The interaction analysis feedback is more effective than unstructured generalizing remarks.

Over time, the concept microteaching, the conceptual and theoretical bases became an important component of competency-based teacher education; the concept of technical skills of teaching has encouraged also teacher educators in the area of physical education to adopt an analytic approach to teaching effectiveness.

Systematic observation and interaction analysis of teaching aims at identifying and describing of teaching patterns (or models), of teaching events, and skills and evaluating their impact on students' achievement. In evaluation research connected to teacher education, the teacher and teaching behavior modification are of central concern and therefore an accurate description of instructional process is the main issue. Observation method is the classical means to gather data in ROTE-PE studies and still used in the 1990's research. In connection with physical education teacher training, many strategies and techniques have been used without any systematic planning and control of the program in the 1970s without estimating the validity of their basic components, taxonomies were used and/or modified as well as target behavior without any analysis of their impact on students teaching behavior. The students' time utilization was found to be clearly one of the most powerful indicators of teacher effectiveness in physical education and studied also in microteaching setting.

The experimental studies connected to the effectiveness of different training methods have produced mixed results. The descriptive-analytic multidimensional designs were found to be useful for long-term program evaluation research. Teacher education is a complex area and starting to construct a minicourse course package needs to follow the steps used by Borg and his colleagues (1970): the approach of guided inquiry into teaching where research and development as one bridged the gap between research and practice.

In the late 1980s and in 1990s, clear evolution in ROTE-PE was evident. The use of multidimensional study approaches enlarged the perspective. Data were collected from many different sources. Also contextualizing, the use of ethnographic interpretative approaches, was recognized as important for extending the perspective and to help to understand physical education from the point of view of teachers and students in a frame of programs' contextual settings and communities.

11 PHASE I: PILOT STUDY, EVALUATION OF CURRICULA

11.1 Introduction

11.1.1 Background and purpose

The pilot study, which has been dealt with also in earlier publications (Heinilä 1977b, 1987, 1997), is concerned with the development and evaluation of teacher education programs in terms of process criteria (changes in teacher's verbal and nonverbal behavior). This evaluation was undertaken as a comparison between effects of two microteaching settings which differed with regard to (1) modelling, (2) sequencing of teaching, (3) timing, (4) number of students, and (5) number of reteachings. The evaluation is primarily descriptive and judgemental and its purpose is to indicate the degree of congruence between what is intended and what actually occurs.

This pilot study describes mainly the educational intentions at the curriculum level, the degree of realization at the observation level and students' evaluation at the product level by means of an experimental set-up. It also describes students' ratings of the program based on their own experiences in two different settings. The model in Figure 24 illustrates the frame of reference and strategy used in this pilot study.

The assumptions based on literature surveys presented in the introduction of the Section II, Chapter 1 and Chapter 4) are the basis for problem setting in this study. The instructional process in microteaching was realized with the help of the conceptual framework and the activity form paradigm of PEIAC/LH-75, II presented in the first section, (see Chapter 6, Fig. 20, p. 170 and Figure 24, p. 195).

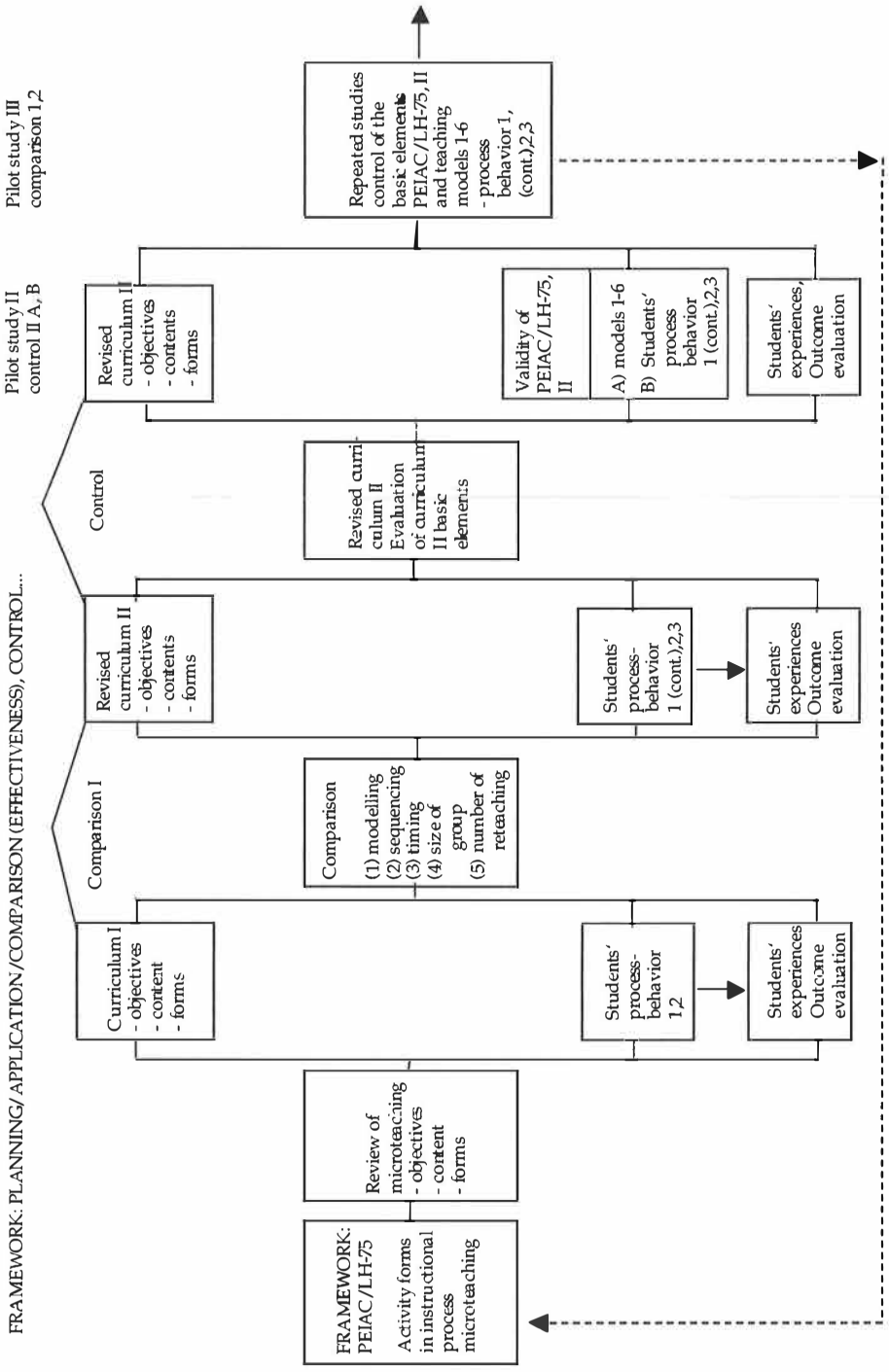


FIGURE 24 A model for descriptive and judgemental curriculum evaluation in terms of process criteria (Heinilä 1977b)

11.2 Problem setting

The main elements of the curriculum package are described in appendix 4 and later are phase 6 pilot study A-D. They can be briefly described as follows:

Objectives
Knowledge and mastery as well as cognitive understanding of characteristics of indirect verbal and nonverbal teaching behavior in P.E. as defined in author's adaptation of Flanders Interaction Analysis Systems, PEIAC/LH-75, II. (Heinilä 1977b) Table 37 and 38, p. 201-202.

Contents
Lectures 15 hours (theoretical background of selected teaching models (1-6); instrument of observation PEIAC/LH-75, II model demonstrations and measurements are used. (See pilot study IIA p. 199-200)

Form of teaching and organization
Practice 30 hours: information, teach one (control 5 min) planning of microlesson one (10 min), teach one, videotape replay, self-observation, analysis, evaluation and discussion; replanning, reteaching (10 min), videotape replay self-observation, analysis, evaluation and comparison of microlessons one and two, summative evaluation. (Appendix 4, see pilot study II, p. 209-210)

During the microlessons, the members of the course group (n=5-10) served as pupils for their classmates, then observed the lessons given by all other students on videotape, and took part in the analysis and discussions.

Material
Handout
Task plan, timing, frame factors (teaching model, subject area, pupils' age level, competence), lesson plan form (Appendix 4.1.2).
Instrument of observation, coding sheet (Table 36 and Appendix 4.1.4).
Timeline display (Appendix 4.4).
Model demonstration videotapes.
Questionnaire for students' course evaluation (Appendix 5.1).
Questionnaire for students' "ideal" P.E. teacher expectations study (App. 8.1).

Evaluation
Formative evaluation (1) planning of microlesson and revised lesson 2; (2) process behavior, model demonstration, (lesson 1 and 2) and coding of own and peers videotaped lessons and analysis of the data. Summative evaluation: test of handout and lectures (1-3 points), practice (1-3 points). Total points (1-3), in degree requirements. Student's evaluation of microteaching course.

11.2.1 Research task

The questions to be answered by the pilot study concerned the form, contents and timing of the course in microteaching. They included:

1. How should students be informed of target behavior?
2. In what way should theory be incorporated into the teaching program?
3. How many microlessons and reteachings are needed?
4. How long should microlessons last?
5. How much time is needed for the analysis of feedback after self-observation?
6. What is the optimum number of pupils in microlessons?
7. Does the constructed observation instrument, PEIAC/LH-75, II facilitate model learning and are students able to observe and evaluate their own and others teaching behavior by means of it?
8. How should the course be placed in the total educational program of P.E. teacher candidates?

(See Figure 24, comparison I, p. 195)

11.3 Methods

11.3.1 Design

The research design was the comparison of two versions of teaching programs and the evaluation of the effect of revisions. It was assumed that the level of program realization and learning outcomes, the revised program (1976) would be more effective than the earlier version (1974).

	1974	1976
(1) information about target behavior	written	written and audio-visual
(2) timing of theory instructions	during the course	during the first third of the course
(3) number of "pupils" in microlessons	4	9
(4) length of microlessons	5 min.	10 min.
(5) number of microlessons	2	3 (of which 1 was used or information and control measurements)

It would have been possible to derive an experimental design on the basis of the above assumptions for studying the effects of different components. In this exploratory study, realized in natural setting, it was decided to aim at obtaining more global descriptive data. The questions concerning the effectiveness of the revised program were based on the following assumptions:

a) At the level of program realization, there are statistically significant differences between the teaching behaviors during microlessons 1 and

2 in the two groups (1974 and 1976) in terms of proportional distribution of time in different categories of the PEIAC/LH-75, II instrument or in the selected indices formed on the basis of them (1, 2, 3, 5 and 7) in direction desired.

b) At the level of program realization and learning outcomes, there are statistically significant differences between the 1974 and 1976 groups in students ratings that concern (1) information about target behavior, (2) timing of theory instruction, (3) number of "pupils" in microlessons, (4) length of microlessons, and (5) number of microlessons as expected.

11.3.2 Subjects

The study focused on the congruence between the objectives of the two-microteaching courses held in 1974 and 1976 and the actual outcomes in terms of process criteria. The subjects were the female and male third year students (1974, n=27; 1976, n=74) at the Faculty of Health and Physical Education at the University of Jyväskylä, participants of microteaching courses.

11.3.3 Procedures

Two measuring instruments were used in the study: (1) the PEIAC/LH-75, II (Heinilä 1977b) was used in observation and coding at student's video recorded microlessons 17500 6 sec time units. The verbal and nonverbal behavior of the teacher and students (Table 37). The microlessons 1 and 2 were recorded by means of CCTV-System of the Faculty (Appendix 2). There was a manually controlled camera on site and a camera manipulated from the control room. The wireless throat microphone used by the teacher recorded the teacher's voice and partly also the voices of the "pupils". The material was processed statistically at the computer centre of the University of Jyväskylä with Honeywell 1944 time sharing-system and HYLPS statistical programme package UPLI-FH.TV-UNIVAC1108 and since 1983 SPSS.

TABLE 37 Categories of modified PEIAC/LH-75[1] II (1)

Cluster 1		
Teacher talk	Response	1. Praises, encourages, accepts the affective tone of a pupil 2. Gives corrective feedback, directs, clarifies, answers pupil's questions 3. Makes use of the ideas and movement patterns suggested by a pupil or group of pupils: 3.2 Clarifies, expands, builds questions and movement initiations on the ideas expressed by a pupil 3.3 Summarized pupil's ideas or movement patterns, asks a pupil to demonstrate 3.4 Compares the ideas or movement patterns expressed by one pupil to those of another or to those given, repeats pupil's ideas, asks a pupil to demonstrate
		4 Asks questions, initiates, terminates activity: 4.1 Asks questions requiring narrow answers, initiates short-term activity, terminates act. 4.2 Broad, open questions which clearly permit choice in ways of answering and moving
	Initiation	5 Content emphasis: 5.1 Presents information, opinions, demonstrates movement patterns, makes a pupil demonstrate 5.2. Organizes pupils, material, division of labour and responsibility 6 Gives directions, commands during activity (pupils expected to comply) 7 Criticizes pupil behavior, rejects movement pattern, justifies authority
Pupil talk		8. Pupil answers question, made by the teacher 9. Pupil initiates speech, asks for instructions, expresses own ideas or movement patterns
Silent teacher activity and other		10. Teacher follows pupil's activity-silent guidance 11. Teacher's silent participation in movement act. 12. Confused situation, uproar
Cluster II Pupils collective movement behavior		1. Pupils collectively passive 2. Pupils collectively active

[1]PEIAC LH/75, Heinilä 1977b, 15

TABLE 38 Indices of PEIAC LH/75 II and their calculation

DEFINITION OF SELECTED INDICES APPEARING IN CONNECTION WITH PEIAC/LH-75 II

INDICES

1. Percent teacher talk (TT) $= \frac{\text{categories } 1,2,31,32,33,41,42,51,52,6,7}{\text{row totals cluster 1}} \cdot 100$

2. Percent pupil talk (PT) $= \frac{\text{categories } 8,9}{\text{row totals cluster 1}} \cdot 100$

4. Teacher's silent guidance and silent participation in movement activity ratio (TSGPR) $= \frac{\text{categories } 10,11,12}{\text{categories } 1,2,31,32,33,41,42,51,52,6,7,10,11,12} \cdot 100$

5. Teacher response ratio (TRR) $= \frac{\text{categories } 1,2,31,32,33}{\text{categories } 1,2,31,32,33,6,7} \cdot 100$

6. Teacher's corrected response ratio (TRRR) $= \frac{\text{categories } 1,2,31,32,33,41,42}{\text{categories } 1,2,31,32,33,6,7,51,52} \cdot 100$

7. Content emphasis ratio (CCR) $= \frac{\text{categories } 41,42,51,52}{\text{row totals cluster 1}} \cdot 100$

Reliability of observation

Reliability was estimated by means of Scotti's *pi* coefficients between the codings by the researcher and the outside trained observer. When the reliability index of a sample of two 60-minute videotaped coding sessions reached the level of .78, the observation of the research data was started. Reliability coefficients are reported in the table 39 (Heinilä 1977b).

TABLE 39 Means of Scott's coefficients for Inter-coder agreement, within-coder, constancy, between-coder constancy by cluster (I,II) and by occasion (T_1, T_2) in microteaching observations (n=11 microlessons)

	CLUSTER I Teacher Talk Pupil Talk Silent Activity		CLUSTER II Pupils' Collective Movement Behavior Activity/Passivity	
	$\bar{x}$	SD	$\bar{x}$	SD
INTER-CODER AGREEMENT				
Videotape Recording T_1	.81	.08	.84	.12
Videotape Recording T_2	.72	.06	.83	.09
WITHIN-CODER CONSTANCY				
A T_1. T_2	.82	.07	.90	.10
B T_2. T_1	.75	.08	.86	.08
BETWEEN-CODER CONSTANCY				
T_1. T_2 A-B	.71	.07	.81	.11
T_2. T_1 B-A	.79	.06	.87	.12

T_2 = 2nd observation (2 months after T_1)

Questionnaire

A questionnaire was used to classify the student teachers' reactions toward the course program and the form of its realization (Appendix 5.1). The questionnaire consisted of 58 items presented as positive and negative statements to which the experimental population were to react by choosing one of five steps on a scale ranking from "disagreement" (step 1) to "uncertain" (step 3) to "complete agreement" (step 5). A modified three-step scale tested distributions of frequencies with Chi Square test (Appendix 5.2). The reliability of the questionnaire, in terms of the homogeneity of variance (Cronbach's alpha), was computed on the item-test correlations and of seven varimax factors sum scores across four populations (n=197). It ranged from .50 to .92 (Heinilä 1988).

11.4 Results

11.4.1 Students' teaching behavior

The data presented in Table 40 show the results of a one-way analysis of variance (ANOVA) for the percentages of category distributions in two clusters and some selected indices based on them indicating the behaviour used by students (n=27 and n=74) participating in two different versions of the studied practice teaching program. The data were based on the marks of a reliable observer who coded the events of the videotaped microlessons One and Two, given by students at one-week intervals. Double coding by using PEIAC/LH-75, II observation instrument was done at six-second intervals.

In the comparison, the testing of hypothesis of the effects of revisions made to program construct on student criterion behavior, statistically significant F-values were obtained in 10 out of 16 analyses of category distributions and in 4 out of 5 analyses of indices (Table 40). This would indicate that the assumptions concerning revision made was supported with regard to these process variables. These results indicated that the revised course program differed clearly from the first version on the level of realization. On the basis of the comparison between selected indices (presented in the figure 25), the differences can be described in the following fashion: there was a difference (a) in the percentage of teacher talk (TT) (from 76%, 1974, to 68%, 1976), (b) in the "silent teachers" didactic activities (TSGPR) (19% to 29%), (c) in the amount of teacher response behavior (TRR) (ID ratio index) (59% to 74%), and (d) in the proportion of the content emphasis (CCR) (47% to 42%).

Furthermore, an examination of the F-values and t-values of statistically significant category distribution differences showed that the behavior of the student teachers of the revised course differed in the following ways: (a) the teacher gave less corrective feedback and answered more to pupil's questions, (b) made much more use of pupil's ideas and movement themes by extending

(cat. 3.1.), summarizing (cat. 3.2.) and comparing them (cat. 3.3.), (c) the teacher asked fewer questions which pupils were expected to answer in a given way or initiated and terminated movement activity (cat. 4.1.), (d) the teacher asked more broad and open questions demanding a higher level of thinking which clearly permitted choices in ways of answering and moving (cat. 4.2.), (e) the teacher presented and demonstrated information and his/her own opinions less (cat. 5.1.), (f) the amount of teacher ordering and direction during movements (cat. 6) decreased as well as (g) the amount of criticism and rejection of pupil behavior or movement pattern (cat. 7), and (h) pupil-initiated talk decreased (cat. 9), whereas (i) the amount of teacher's silent didactic activities (cat. 10-12) increased. It is worth mentioning that these changes were not observed to have influenced pupil's collective movement activity (MET), which was just over 50% of the time in the microlessons of both groups as prescripted in curriculum objectives.

The two courses differed quite clearly with regard to the above-mentioned respects in terms of both the first and second microlessons, whereas differences between the two lessons within courses were small.

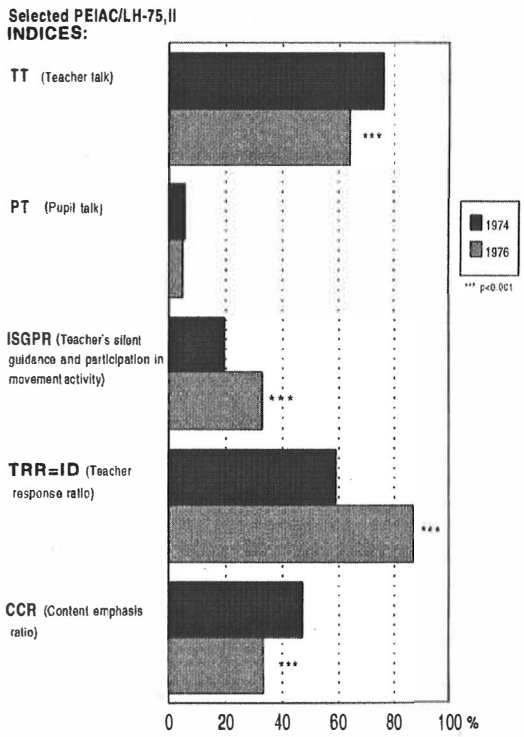


FIGURE 25 Comparison of curriculum groups 1974 and 1976 on the percentage of index means in microlessons 1 and 2.

TABLE 40 Comparison of the curriculum groups 1974 and 1976 on the percentages of behavior used microlessons 1 and 2; ANOVA and t-test computed by categories of Clusters I and II and by selected indices based on row totals

Categories and indices	Curriculum 1974 1st Lesson (N=27) x	SD	2nd Lesson (N=27) x	SD	Total (N=54) x	SD	Diff. 1-2 df=52 t	Curriculum 1976 1st Lesson (N=74) x	SD	2nd Lesson (N=74) x	SD	Total (N=148) x	SD	Diff. 1-2 df=146 t	Diff. 74-76 Diff. 1-1 df=99 t	Diff. 2-2 df=99 t	F df=1 df=199 t
I Cluster																	
1.	5.8	3.0	4.8	3.6	5.3	3.3	-1.07	5.1	3.9	4.6	3.4	4.9	3.7	-.75	-.83	-.22	0.57
2.	7.2	6.1	7.0	6.5	7.1	6.2	-.13	5.1	4.7	5.1	4.8	5.1	4.8	-.02	-1.80	1.56	5.63*
3.1.	2.1	3.5	3.7	7.1	2.9	5.7	1.01	12.7	6.7	13.6	6.1	13.2	6.4	.97	7.72***	6.92***	107.31***
3.2.	0.1	0.5	0.1	0.5	0.2	0.5	0.0	1.4	1.6	2.0	2.5	1.7	2.1	1.58	4.07**	3.80**	29.35***
3.3.	0.8	2.4	0.4	1.2	0.6	1.9	-.71	1.4	2.3	1.9	2.8	1.6	2.6	1.27	1.06	2.61*	6.88**
4.1.	13.0	7.0	13.1	5.0	13.0	6.0	.09	8.6	4.3	8.5	3.8	8.5	4.1	-.20	-3.76***	-4.96***	36.73***
4.2.	0.4	1.0	1.0	2.8	0.7	2.1	1.02	1.9	1.4	1.9	1.3	1.9	1.4	-.18	5.25***	2.23*	23.66***
5.1.	29.6	11.7	27.6	12.2	28.6	11.9	-.64	19.4	8.2	17.1	7.2	18.2	7.8	-1.79	-4.94***	-5.29***	52.52***
5.2.	5.6	4.4	4.6	3.4	5.1	3.9	-.90	5.1	3.4	4.4	3.0	4.7	3.2	-1.30	-.58	-.28	.39
6.	9.9	9.7	10.3	12.3	10.1	11.0	.12	2.9	3.4	3.2	3.4	3.0	3.4	.54	-5.45***	-4.54***	48.82***
7.	2.7	3.1	2.4	3.6	2.6	2.4	-.40	0.9	1.1	1.0	1.4	0.9	1.3	.33	-4.36***	-2.82**	24.85***
8.	3.4	3.8	2.5	2.4	3.0	3.2	-1.04	2.8	2.3	3.4	2.7	3.1	2.5	1.60	-1.05	1.52	0.08
9.	2.4	3.2	2.7	2.7	2.6	2.9	.46	1.4	1.6	1.6	1.7	1.6	1.7	0.0	-1.64	-2.57*	8.74**
10-12.	16.9	14.4	19.8	15.0	18.3	14.6	.72	31.2	12.1	31.7	12.6	31.5	12.3	.23	5.00***	4.00***	40.47***
II Cluster																	
1.	50.2	14.6	47.5	13.0	48.9	13.7	-.73	48.9	12.5	48.2	12.2	48.6	12.3	-.34	-.45	.26	.48
2.	49.8	14.6	52.5	13.0	51.2	13.7	.73	51.1	12.5	51.8	12.2	51.5	12.3	.34	.45	-.26	.48
Indices																	
1. (TT)	77.3	13.9	75.0	14.6	76.1	14.2	-.59	64.4	11.0	63.3	11.1	63.9	11.0	-.61	-4.81***	-4.28***	41.51***
2. (PT)	5.8	5.2	5.3	3.8	5.5	4.5	-.42	4.3	2.8	5.0	3.2	4.7	3.0	1.32	-1.79	-.36	2.42
4. (TSGPR)	17.7	15.0	20.8	15.5	19.3	15.2	.72	32.5	12.2	33.1	12.7	32.8	12.4	.33	5.04***	4.08***	41.53***
5. (TRR)	54.4	21.8	60.2	30.8	59.3	26.4	.25	87.4	12.2	86.7	11.6	87.0	11.9	-.33	8.39***	6.30***	104.28***
7. (CCR)	48.5	12.9	46.2	11.5	47.4	12.2	-.69	35.0	10.7	31.9	10.1	33.4	10.5	-1.82	-5.31***	-6.09***	64.81***

*, **, *** =p<.05, p<.01, p<.001 respectively

In summary, it may be stated that at the level of realization of the course program, the group whose program had been revised with regard to (1) information about target behavior, (2) timing of theory instruction, (3) number of pupils in the microlessons, (4) length of microlessons, and (5) number of microlessons, displayed more indirect behavior which had been set as a goal. The teacher offered the pupils more opportunities to create ideas and solve problems, was more inclined to observe pupil responses, and took advantage of these responses in the topic treatment. Pupil-initiated talk did not increase. However, this may be due to the type of pupils who may have been less inclined to "role playing" or the teacher may have directed his main attention to movement ideas and activity.

11.4.2 Student ratings of the microteaching course

From questionnaires filled out by the students (n=121), data were obtained on student reactions to revisions made. The significance between the percentage distributions of statements classified to three groups divergent opinion, uncertain, total agreement was tested by the Chi Square test. (Appendix 5.2) There were statistically significant differences between the answers given by the students of the two courses. Contrary to the students of the first course, the students attending the revised course were of the opinion that the course could well be placed in the third year study program, not before (15) from 15%, 1974, to 60%, 1976, agree, ($x^2 = 25.7$, $p < .001$). The course did not, in the opinion of the students of the revised course, overlap with other teaching (11) 69% to 89% agree, ($x^2 = 7.75$, $p < .05$) and they were more interested in the theory lessons (26) from 21% to 41% agree, ($x^2 = 7.95$, $p < .05$).

The students of the revised course were more satisfied with the amount of use of audiovisual material (19) from 21% to 47% agree, ($x^2 = 15.47$, $p < .001$) but they still wanted more. The students of the revised course thought that the time available for exercises was not sufficient, however, they were more satisfied with the time arrangement than the students of the first course (34) 73% to 60% disagree, ($x^2 = 6.15$, $p < .05$). The students of the revised course were more satisfied with the use of peer students as pupils than the student of first course, (10) from 18% to 56% agree ($x^2 = 16.9$, $p < .001$).

The students of both courses reported that the demonstrations of lecture and teaching models would not be sufficient without having to participate in exercises (50) from 90% to 92% ($x^2 = 8.01$, $p < .05$). The students of the revised course were less satisfied with the selection of exercises ((9) from 18%, 1974, to 41%, 1976 ($x^2 = 9.25$, $p < .01$) but they thought that the exercises were sufficiently varied (16) from 27% to 53% agree ($x^2 = 12.9$, $p < .01$).

The students attending the revised course were more often of the opinion that the course had opened a new outlook (46) from 50% to 66% agree ($x^2 = 6.63$, $p < .001$) and the organization of the course was judged to be better (52) from 33% to 69% agree, ($x^2 = 4.72$, $p < .001$). The students of the revised course considered to have learned better than the students of the first course to discriminate between teaching patterns in observing and coding feedback (53) 45% to 82% agree ($x^2 = 19.00$, $p < .001$).

In addition, the students of both courses were very satisfied with lecture handouts (28) in both 1974 and 1976, 92% agree; thought that demonstration tasks had been well selected (47) from 70% to 62% agree and lectures and demonstrations were well coordinated (56) from 50% to 69% agree. The students reported that the course had been useful (35) from 77% to 87% agree) and that they intended to use in their future practical teaching the teaching patterns they had learned (58) from 79% to 82% agree. They also thought that their views of teaching behavior had broadened (57) from 77% to 88% agree and that during the course they had become aware of errors and weaknesses in their teaching behavior (59) from 68% to 78% agree.

11.5 Summary and conclusions

Two versions of a practice teaching program have been described and compared. The congruence between the intended and actual outcomes was examined in order to draw conclusions about the rationale of component revision and to provide some basis for the placement of the different modifications of the course in the P.E. teacher education program.

The congruence between objectives, which were identical in both programs, and the degree of their realization, was improved in the revised program judging from observation of the student's teaching behavior and their ratings of the courses. The revised program, which included written and videotaped materials, instruction of theory during the early part of the course, and microlessons with nine students and lasting 10 minutes, proved more effective than the original. The students applied better patterns of indirect teaching and were aware of and understood better their theoretical background. The differences between the first and second microlessons of both courses were not statistically significant in terms of any variables. It follows that the number of reteachings should be carefully considered as well as developing their contents and the gradual increasing of level of difficulty.

The instrument of interaction analysis PEIAC/LH-75, II modification used in the courses was based on the theory of Flanders (1965, 1970) and his FIAC system and on empirical study of physical education teaching and framework (Heinilä 1974, 1977b). It proved feasible both from the point of view of research and of teaching. It appeared to facilitate the operationalization, information, evaluation and measurement of intended behavior, code patterns. However, the construct validity and sensitivity of the basic elements of the program needs to be investigated and estimated more closely. (This has been done in the next phase of the study.)

12 PHASE II: THE VALIDATION OF THE BASIC ELEMENTS OF THE MICROTEACHING PROGRAM

12.1 Introduction

12.1.1 Background and purpose

The development of new programs for teacher education presupposes the design of the teaching strategy models, and the evaluation and assessment of their basic elements. In the following Multiple Baseline design II illustrated in Figure 26, the causal relationships between the goals, directed interventions, assessment of change of students learning gain and summative evaluation in revised PETE-study unit program of didactic observation and microteaching are presented.

The aim of this section is to report on the assessment and evaluation of the following critical redesigned program components and processes:

A. PEIAC/LH-75 II observation instrument and observation of non-directive teaching models: validity and sensitivity
B. PEIAC/LH-75 II observation instruments (construct validity, factor structures)
C. Students' entry teaching skills, rating scale (reliability, stability, validity and sensitivity)
D. Students' entry attitudes, a questionnaire concerning expectations of an "ideal" P.E. teacher's characteristics, (factor structure, reliability and validity)

12.1.2 Multiple baseline design (phase II): intervention strategy

The objectives were to investigate and describe the causal relationship between the intervention and change in students' knowledge, mastery and cognitive understanding of the characteristics of teacher-student interaction as defined by Heinilä's adaptation of Flanders' Interaction Analysis System (Flanders 1970), PEIAC/LH-75 and PEIAC/LH-75 II (Heinilä 1977a, 1977b, 1987) in pre-service

teacher training course of didactic observation and microteaching (62 h/2 study weeks) at the University of Jyväskylä, Finland (1974-1991)

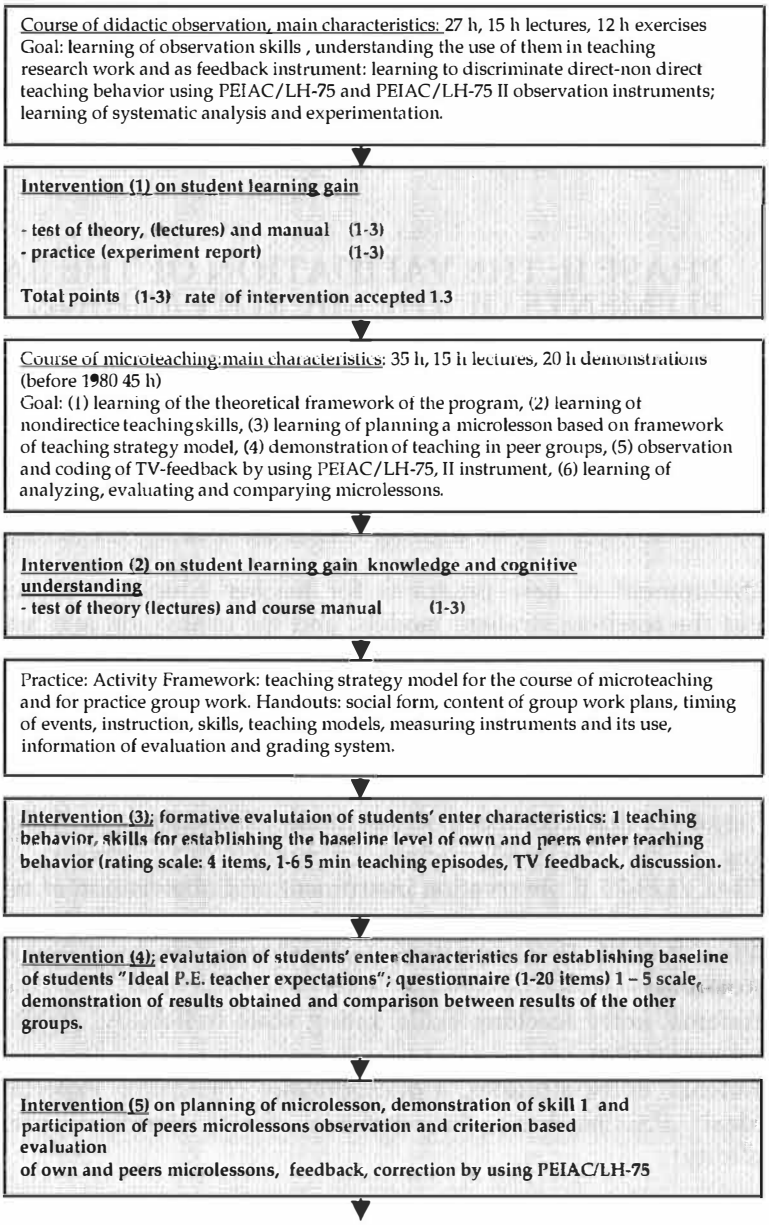


FIGURE 26 (continues)

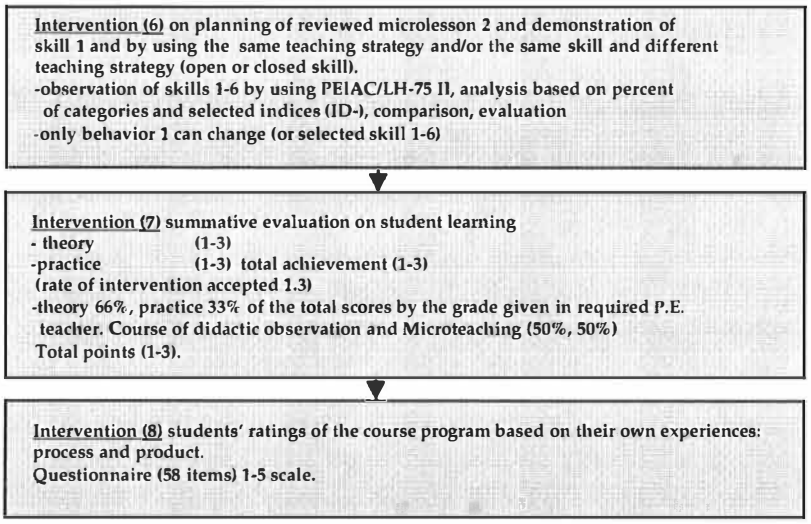


FIGURE 26 Components of the program teaching strategy model, based on curriculum framework by using PEIAC/LH-75 II system (Heinilä 1977b)

12.2 Pilot study II A: The validation of an observation system: a multivariate approach

12.2.1 Introduction

(1) Background and purpose

The basic goal of this particular investigation was to design and validate an instrument, which would take into account the clear specification of task to be learned and to assess students' engagement with the criterion task as determined in the curriculum.

The objectives of the training program included knowledge and mastery as well as a cognitive understanding of the characteristics of indirect verbal and non-verbal behavior as defined in Heinilä's adaptation of Flanders' Interaction Analysing system PEIAC/LH-75 teaching (Heinilä 1977b, 1990) in physical education classes. This pilot study was based on the revised microteaching course whose basic objective was to familiarize the students with general Flanders-originated ideas of indirect verbal and non-verbal teaching behavior (Flanders, 1965, 1966, 1970, Heinilä, 1977b, 1990) and with more specific models or patterns of indirect teaching, which are operationalizations of indirect teaching (for a description of these models (see the next pages 210-212). Students then gave two 10 min. microlessons between one-week intervals demonstrating their learning of the ideas. Microlessons were video recorded

and observed from videotape by students using a modified version of the developed observations systems (PEIAC/LH-75 II, Heinilä 1977b).

(2) The modified observation instrument PEIAC/LH-75 II subscription

When modifying a category system, the word "subscripting" means dividing a single category into additional subcategories. The term was first used in the early development of coding at the University of Minnesota Laboratory for Research in Social Relations, and at Michigan by Harold Anderson (Amidon & Hough 1967). The utility of subscription depends on the close correspondence between the problem elements, concepts and category definitions.

A number of researchers have subscripted the Flanders FIAC system in order to increase the discriminations amount the statements. For example, the subscripts developed by Galloway, Honigman and others can be found in the anthology of category systems by Simon and Boyer (1970). Other examples in the area of physical education are the works of Cheffers, Mancuso, Dougherty, Gasson and Underwood (see Chapter 2, Section I).

Subscripting can provide additional data. The kind of additional data that would be most useful depends on the purpose of the observation. The relationship between the purpose and procedure might be clarified as by Flanders' 22-category system which was constructed by subdividing the original 10-category system for combining process and a cognitive orientation. It was designed primarily to subdivide category 3 with additional subscripts suggested by members of the trained (3 weeks of 6 hours per week) six observers team for which eighteen of nineteen reliability checks produced a Scott Pi coefficient with the median at 0.79 (Flanders 1970, 140 - 141).

The two clusters 18-category system presented in the Table 37 is designed primarily to subdivide three categories of the first cluster of PEIAC/LH-75 for combining process and a cognitive orientation in microteaching. As seen in Table 41, in PEIAC/LH-75 II the first cluster (teacher talk, pupil talk, silent teacher activity) is enlarged, because three categories (3, 4 and 5) seemed to be too narrow for coding teaching models. The second cluster, (movement) activity was too large to be used as such in this connection, however it was important to control time of students motor engagement (MET). The categories of the third cluster (social form) were used in the pre-interactive phase, by organizing the group activity and, planning of lessons as frame factors - (teaching strategy model, Figure 27, p. 213). In the first cluster of the modified instrument, subcategories were added to category three (31, 32, 33) to category four (41, 42) and five (51, 52). In the second cluster only the two main categories, pupil's collective activity/passivity are used (MET). Double coding is made at six-second time intervals. The coding sheet and instructions is presented in Appendix 4.5.

TABLE 41 The classification system for observing the microlessons (cluster I of PEIAC/LH-75 II (Heinilä, 1977b)

Teacher talk	
	Response
1.	Praises, encourages, accepts the feeling tone of a pupil
2.	Gives corrective feedback, directs, clarifies, answers pupil's questions
3.1	Makes use of the ideas and movement patterns suggested by a pupil: clarifies, expands, builds questions and movement initiations on the ideas expressed by a pupil
3.2	Summarizes pupil's ideas or movement patterns, asks a pupil to demonstrate
3.3	Compares the ideas or movement patterns expressed by one pupil to those of another or to those given, repeats pupil's ideas, asks a pupil to demonstrate
	Question
4.1	Asks questions requiring narrow answers, initiates short-term activity, terminates activity
4.2	Makes questions requiring higher level of thinking or activity
	Initiation
5.1	Presents information, opinions, demonstrates movement patterns, makes a pupil demonstrate
5.2	Organizes pupils, material, division of labour and responsibility
6.	Gives directions, commands during activity (pupils expected to comply)
6.	Criticizes pupil behaviour, rejects movement pattern, justifies authority
Pupil talk	
	Response Initiation
8.	Pupil answers question made by the teacher
9.	Pupil initiates speech, asks for instructions, expresses own ideas of movements
Other	
	Silence, confused situation
(10)	Teacher follows pupil's activity, silent guidance
(11)	Teacher's silent participation in movement activity
(12)	Confused situation, uproar
The decision on classification is made on the basis of the didactic function of the activity	

(3) The indirect teaching models

The teaching model is a phenomenon constructed from different correlative elements. In interaction analysis component-skills refer to the sequence of teacher-pupil interaction and are called "patterns of teaching". A teaching pattern or model is related to the process behavior and can be identified with the help of the process analysis technique (Flanders, 1970). The systematic observation method enables the quantification and measurement of the features of the teaching-learning process, such as teacher initiation - pupil initiation, and direct - nondirect as well as specific direct and nondirect features as teaching models. In teacher training some measure of performance is used to control the process itself. *The performance criterion is improvement,* and measures of

improvement should be used in making decisions about the feedback available to teachers (Flanders 1987, 20).

The indirect teaching models used in the course of microteaching were (Heinilä 1977b):

1. Teacher initiatives based on pupil responses
The P.E. teacher has to be able to make use of pupils' earlier performance or initiatives by making questions and suggestions related to them or by making the pupil demonstrate his performance. The teacher must then clarify essential points.

2. Summarizing model
The P.E. teacher has to be able to summarize what pupils have done or said and then proceed to the next logical stage by making use of the summary. He can also make pupils demonstrate the functional solutions of the sub-stage and describe them verbally. This is effective reinforcement of pupils' initiative.

3. Comparison model
The P.E. teacher has to be able to observe and compare pupils' movements or their previous ideas to other pupils' movements or given task requirements. In this way the teacher can help pupils to solve problems and guide them to identify key ideas while showing or giving the pupils the impression that they solved the problems on their own. This kind of teacher activity, in which pupils' performance is informed or described to other pupils serves to reinforce their initiative and independent behavior.

4. Model of guiding feedback
The P.E. teacher has to be able to give guiding feedback to the whole class, smaller groups and individual pupils. The giving of feedback presupposes exact definition of objectives and tasks. The teacher has to be able to give feedback wisely, in a variety of ways and giving reasons for his statements. The use of guiding feedback is common in physical education. For instance, in the teaching of some "closed" motor skill (in given circumstances and restricted) it has a decisive role. The role of guiding feedback is to help a pupil to become aware of his performance and to find solutions to problems concerning e.g. movement paths, timing, use of power or space. Giving guiding feedback with statement of reasons for it will help to promote independence. The teacher has tried to see the pupil as a person with whom things can be discussed and planned before decisions are made. The pupil can thus be guided towards a goal, which he understands and accepts.

5. Model of reinforcement and extinction
The P.E. teacher has to be able to observe - to watch and listen to - pupil's ideas and movements with a view to organizing them in terms of teaching objectives and to reinforce selectively those ideas and movements, which are on target. The teacher also has to be able to state without hesitation and clearly what is not relevant or useful from the point of view of the teaching objective. Such

responses may be directed to the whole class, to smaller groups or to individual pupils. Praise and reward and criticism may concern pupil's behavior or movements. Praise can be verbal but also symbolic (e.g. smile, applause), similarly rejection. In acting on pupils' conditions the teacher's reasons must be related either to the whole class, to groups of pupils or to individual pupils.

6. Discrimination model
The P.E. teacher has to be able to clarify - verbally and through demonstration -- the logic of classroom discourse and progress. For instance, he can clarify the pupils' degree of freedom of social activities by stating given or accepted directions - customs, norms, rules of the game, etc. This includes the maintenance of a consistent meaning of words, concepts and movements. Accurate concepts aid communication and classroom discipline is improved. It is especially important to help pupils to distinguish between facts, opinions and valuations. This presupposes that the teacher monitors and evaluates the situation.

(4) Properties and activity forms in the paradigm of PEIAC/LH-75 II

Teaching strategy model

Teaching strategy model based on the curriculum framework presented in Section I, Chapter 6 (see p. 170) by using observation instrument PEIAC/LH-75 II (Heinilä, 1977a, 1977b, 1983, 1990) is illustrated in Figure 27, p. 213. It was used in the pre-interactive phase for organizing the group-work activity, with the purpose to frame the students' pre-interactive planning of microlessons and to implement their understanding of the use of different teaching models in contextual variation. Every student of the group was asked to make a lesson plan and to use one model first and to apply it in a certain subject area for pupils of certain age and for more or less skilled pupils. If the student succeeded well in the interactive phase in the first lesson in criterion skill, he or she was asked to use another model or to select another teaching style for the second microlesson. This kind of revision of intervention technique was made, however, after the study degree program reform (1978) at the faculty. In the experimentation presented in this pilot study II A and B, the student were not asked to change the lesson plan, but only to correct or complete it as well as the evaluated teaching process. Thus, the student was training two times the same model. Also he or she was as a pupil (peer) in the other two microlessons of the practice group (three lessons in every session), and observing, TV feedback coding, analysing comparing and participating in evaluation and discussion concerning the other student teachers' microlessons by using different models and different teaching strategy. The students in each group were also asked to play different roles in the microlessons as either pupils of low or high age or more or less skill. The student teacher informed the group about the role expectation before starting the teaching process.

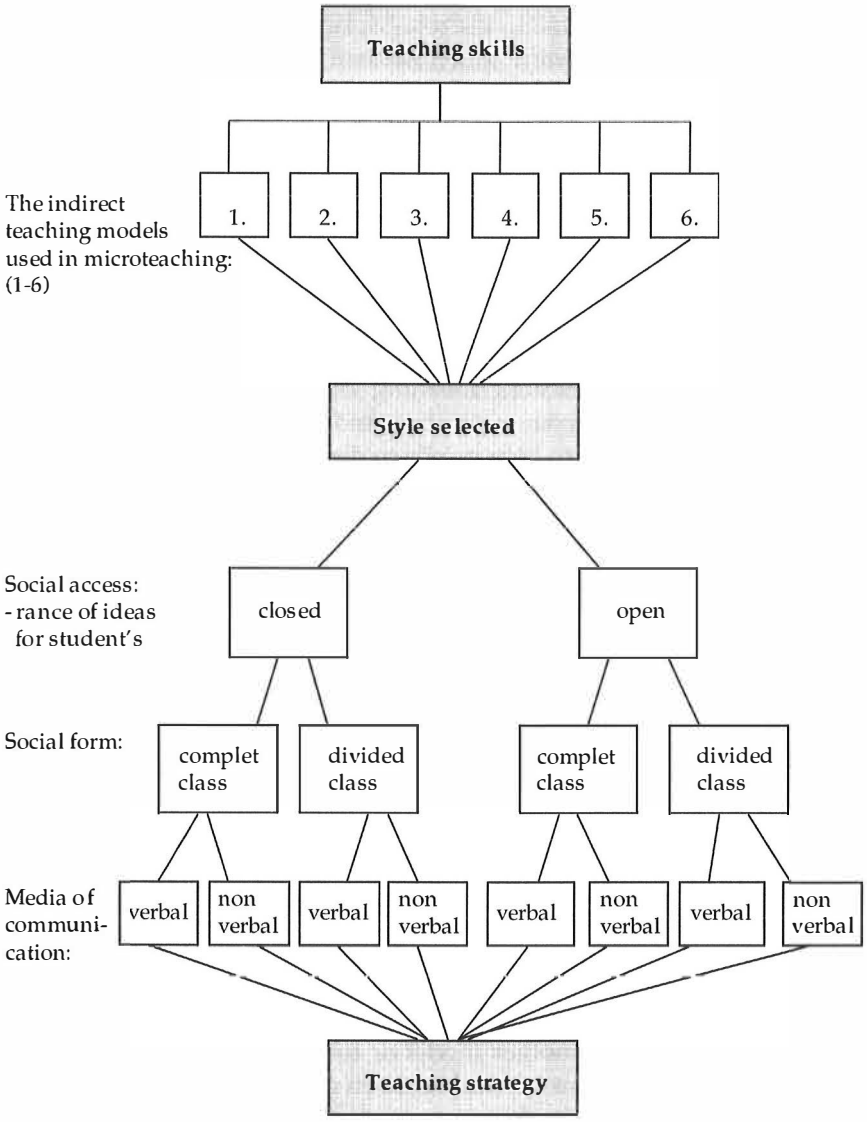


FIGURE 27 Components of the microteaching strategy, (PEIAC/LH-75II) model

12.2.2 Problem setting

The need for adequate concepts

The development of a new program for practice teaching presupposes evaluation and assessment of conceptual integrity and coherence. In modifying a category system and developing teaching models the problem of adequate concepts is a central importance. If they are based on particular theory, it influences the way one sees the world. Thus, the feedback, which occurs in this kind of training system, consists of opinions, ideas, and models representing a certain orientation (see Section I Chapter 5, p. 78).

As stated in Section I, Chapter Three (p. 64), Flanders (1970) was dealing with the problem of validity in terms of models, and stated that the issue of validity in coding depended on whether what was encoded did in fact exist and whether the elements of the original situation were recreated in their proper perspective during the decoding process. Validity, therefore, requires accurate interpretation during both decoding and encoding.

Terms such as "indirectiveness" and the contrast between direct and indirect teaching are however more general than the coded events themselves. The apparent non-quantibility of many variables may be seen as a consequence of the fact that we do not know which are the primary elements of a certain phenomenon as a teaching model, and which are composites of several primary elements. In physical education classes this problem is of central importance. The correspondence between the categories and reality cannot be directly seen, because here the "reality" is internal, not observable. The "pattern" image of the "true structure" must be constructed by reasoning from the concrete, observable facts. Thus, the starting point for constructing teaching models as well as for validation should be the analysis of the correspondence between the categories and "reality".

In this context the categories of the modified 18-category PEIAC/LH-75 II system were assumed to correspond to the teaching models in the following way:

No.	Teaching model	Cluster I Category
1.	Teacher initiations based on pupils responses	3.1
2.	Summarizing model	3.2
3.	Comparison model	3.3
4.	Model of guiding feedback	2., 3.1
5.	Model of reinforcement and extinction	1., 2., 3., 7.
6.	Discrimination model	2., 3., 4., and 5. (Appendix 4.2.3)

12.2.3 Research task

Since the research reported here was part of a larger effort to construct a method for describing interaction in P.E. classes, it is necessary to test empirically the measurement qualities of the developed system. One of the

most important requirements of any measuring device is validity, more specifically construct validity.

Dunkin (1976) has argued that two criteria should be applied in judging the validity of what he calls 'technical skills of teaching'. One has to do with the extent to which the specific aspects of teaching behavior (in the case of the present study, various patterns of indirect teaching behavior) are distinct from other aspects of teaching. Observers should be able to agree on what constitutes the nature of the skill and should be able to identify it when it occurs. The other criterion concerns the extent to which the skills can be shown to improve student learning. This study addresses only the first of Dunkin's concerns.

In a later article Dunkin (1987, 705) suggested that in spite of progress made in identifying teaching skills "it is probably still the case that attempts to specify technical skills of teaching rely more on impressionistic evidence from professional experience than upon systematically obtained evidence from empirical research".

This study is an attempt to contribute in a small way to the empirical evidence research-base that Dunkin calls for.

The aim of this study was:

(a) to find those discriminant functions that best separate the criterion groups from each other, in other words maximize the between-group variance, relative to the within-group variance;
(b) to describe factors connected with the use of the modified category system and predicting the grouping of microlessons, and thus to describe the ability of the instrument to distinguish between model groups "known" to behave differently on the construct under study;
(c) to describe the sensitivity of the observation instrument to make discriminations required for the research problem: combining process and a cognitive orientation.

12.2.4 Method

Multivariate approach to validation in observation research

As stated in the above, this study is concerned with the exploration of the construct validity of a combination of teaching models used in a microteaching course and an observation instrument, both derived from general ideas of Flanders' research paradigm (1965, 1966, 1970).

Construct validity is often determined in an indirect way. The researcher uses a theory to establish a set of hypotheses about how the data should behave. For instance, the researcher predicts certain internal relationships between measured variables: high, intermediate or low correlations. A construct-valid instrument will produce scores that correlate only with those variables with which, on the basis of theory, it should correlate, and the scores of those variables to which it should not be related will not correlate with it (convergent vs. discriminant validity). Similarly, a construct-valid instrument should distinguish between groups that are "known" to behave differently on the construct under study (Campbell & Fiske, 1959).

In order to handle several correlations for validation, a multivariate approach is needed, as was done in earlier stages of this study (Heinilä, 1980). A discriminant analysis was applied in the present study since it may be interpreted (Cooley & Lohnes, 1971) as a special type of factor analysis that extracts orthogonal factors of measurement battery for the specific task of displaying and capitalizing upon differences among criterion group (here six different teaching models).

Design and material

In this validation study, the data were 148 microlessons recorded during the revised microteaching course (n=74), the first and the second microlessons grouped by the subjects teaching-models (1-6) used and observed from video recorded material (see Appendix 6) by using the modified PEIAC/LH-75 II category system (Table 37). Reliability (.78) was estimated by means of Scott's pi coefficients.

A 10-minute microlesson was used as the statistical unit. In this time a particular teaching model might be realized more than once and simultaneously criterion for the pupils' movement activity (MET) was stated 50% of the microlesson time. The lesson plan was based on the Teaching Strategy Model (Figure 27, p. 213).

The primary data consisted of the double codings of six-second intervals in each category in each lesson (74 x 2 x 100). To compensate for the variance in the number of model lessons, percentages were used as the basis for the correlations. (Table 39)

The material was processed statistically at the computer centre of the University of Jyväskylä with the Honeywell 1944 time Sharing System and HYLPS - System programme package UPLI-FH-TV UNIVAC 1108 and since the year 1980 with the Statistical Package for Social Sciences ($SPSS_x$).

12.2.5 Results of discriminant analysis

Content and interpretation of discriminant functions

In this phase five discriminant functions separating the criterion groups "known" to behave differently were found (Table 42). The discrimination of the first three could be considered highly significant at the 1% level, and the fourth almost significant at the 5% level based on the Chi Square tests results computed from Wilks' Lambda. The share of total discrimination for each discriminant function was 48%, 28%, 17%, 6% and 2.2%. The program selected 9 of the 18 categories of the instrument and set them in sequence according to how much they increased the model's discriminating power. Only categories of the first cluster were represented in the model and it was also noted that four of the five subscripted categories were selected to the discriminative model. The structure of the discriminant model was related to the structure of the measuring instrument as seen from RC-values of functions .70, .61, .50, .32 and .21.

TABLE 42 Discriminant function coefficients for the six model groups variables (PEIAC/LH-75 II categories) of the 2nd and 3rd microlesson observations, n=148

Categories No	Function 1	Function 2	Function 3	Function 4	Function 5
3.3. Compares the ideas or movement patterns expresses by one pupil to those of another or to those given, repeats pupil's ideas, asks a pupil to demonstrate	.94	.18	-.16	.11	-.84
2. Gives corrective feedback, directs, clarifies, answers pupil's questions	.18	.77	.63	.35	-.35
3.2. Summarizes pupil's ideas or movement patterns, asks a pupil to demonstrate	-.28	-.35	.54	.22	-.69
10. Teacher follows pupil's activity, silent guidance	.31	.87	.78	.29	.59
5.1. Present information, opinions, demonstrates movement patterns, makes a pupil demonstrate	.24	.13	.90	-.11	-.48
8. Pupil answers questions made by the teacher	.26	.40	.45	.27	-.81
9. Pupil initiates speech asks for instruction, expresses own ideas of movements	-.30	.26	-.41	.53	.35
5.2. Organizes pupils, material, division of labour and responsibility	-.89	.38	-.29	-.56	-.56
6. Gives directions, commands during activity (pupils expected to comply)	-.34	.50	-.18	.33	-.20
Percent of variance explained	48%	28%	17%	6%	2%
Eigenvalue	.95	.55	.34	.11	.41
Wilks' Lambda 0.212					
Chi Square (NDF)	94.1***	61.6***	41.0***	15.1*	6.2 ns
* p<0.05, *** p<0.001					
RC	.70	.60	.50	.32	.21

The interpretation of the contents of discriminant dimensions was based on the following factors: 1) weights of variables on scaled eigen vectors (s), their discrimination power, 2) correlations (r) of discriminant functions with variables selected into the model, and 3) the placement of known groups on the discriminant dimensions.

Content of the discriminant functions

The specific aspect or aspects of teacher response behavior were represented in all discriminant functions, as well as teacher initiation and silent behavior in two functions, and pupil's verbal initiation and response behavior in two functions. From the structure of coefficients of Table 42 and Figure 28 and Figure 29 the four statistically significant functions extracted appeared to measure:

DF I One specific aspect of teacher response behavior, *comparison vs. models* based on several aspects.

DF II One specific aspect of teacher response behavior, *corrective feedback* and silent guidance vs. models based on other aspects of teacher response behavior.

DF III Specific aspects of teacher response behavior, *summarizing, corrective feedback,* specific aspect of teacher initiation, silent guidance and pupil verbal answers to questions made by the teacher vs. other models.

DF IV Specific aspects of teacher response behavior, *corrective feedback, extinction and pupil verbal initiation* vs. other models.

The assumed correspondence of the categories to the six teaching models was shown to be successful for models 2, 3, 4 and 5. In teaching models 1 and 6 the category-reality correspondence was not so clear in this data.

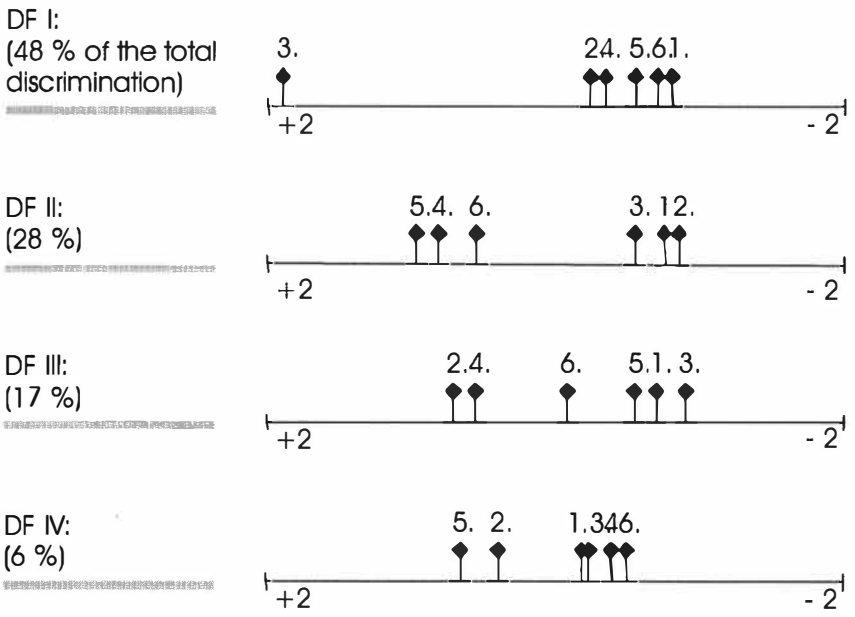
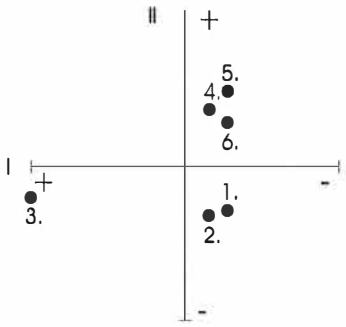
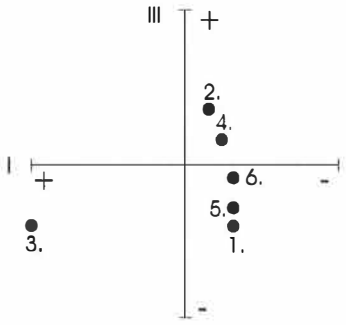


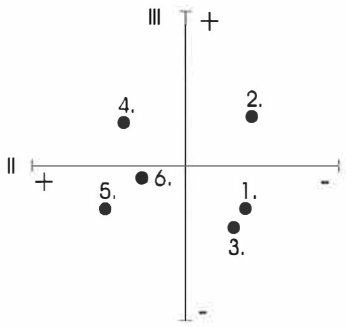
FIGURE 28 Placement of the model groups (1-6) centroids on the discriminant plan formed by discriminant functions I-IV



Discriminant functions I and II



Discriminant functions I and III



Discriminant functions II and III

FIGURE 29 Placement of model groups (1-6) centroids on the discrimination plane on the basis of the means and standard deviations of them in discriminant functions

12.2.5 Phase II A: Discussion and conclusions

Most of the different teaching models could be placed on dimensions formed of four statistically significant (Chi Square $p < .001$, $p < .05$) discriminant functions, reflecting their specific aspects of non-directive teaching. The factors resulting from the modified category system and predicting the grouping of criterion groups (microlessons) could be thus described with PEIAC/LH-75 II system as hypothesized. One aspect of teacher response behavior, "comparison", was quite homogeneous, an "ideal case", whereas the others correlated, as expected, with the other aspects, such as teachers silent guidance, initiation, pupil verbal response and initiation behavior. Two models (1 and 6) might have too large or too complicated contents. The structure of the discriminative model was related to the structure of the measuring instrument (Rc values .70-.60, .50, .32) and produced a clear sequence predicting the grouping of microlessons in accordance with different "known" models used in this study.

Also results obtained in a pilot study conducted in natural setting ($N = 8$ teachers) by Akkanen (1979), by using the PEIAC/LH-75 II system and teaching models designed by Heinilä (1977b) supported results obtained in this validation study.

It was concluded that (a) the two cluster 18-category PEIAC/LH-75 II system for combining process and a cognitive orientation possesses a definite degree of construct validity and objectivity, and that (b) it is sufficiently sensitive to discriminant aspects of direct/nondirect teaching behavior and to a definite degree also aspects of nondirect teaching behavior operationalized as teaching models.

Although the results of this discriminant analysis can only be regarded as tentative due to the nature of the level of the measurement scale and the quality of the sample of observations based on students' training, microlessons 1 and 2, that analysis yielded quite useful information for the development of the instrument, teaching models and intervention technique.

Still one can agree with Dunkin (1987) that there are still many problems with regard to the validity of many specific teaching skills included in teacher educational programs.

12.3 Pilot study II B: Investigation construct validity of an observation instrument: a multivariate approach

12.3.1 Problem setting

As stated in the introduction, the development of programmes of new practice teaching presupposes the controlling and evaluation of their basic elements. The basic goal of this investigation was to design an instrument, which would take into account clear specification of task to be learned and measure students' engagement with the criterion task determined in the curriculum. The curriculum, its objectives, forms and contents were presented in the earlier

chapter (Pilot study I and II A). Since the main element of the revised program (PEIAC/LH-75 II-system) is modified from Flanders' (1965, 1966, 1970) 22-taxonomy and from PEIAC/LH-75 (Heinilä 1977a, 1977b), it was necessary to test the empirical measurement qualities of the taxonomy a in microteaching setting. One of the most important requirements of any measuring device is validity, more specifically construct validity, especially if the results of measurement are meant to be used e.g. in connection with a long-term program evaluation study. The stability of the structural construct needs also to be estimated in contextual variation.

12.3.2 Research task

This study will explore, from the point of view of the Flanders' theory, the interaction in 221 microlessons of the revised course by considering the systematic variance among scores when using the PEIAC/LH-75 II two-cluster category -system on the construct under investigation. In this phase of the study, the aim was to examine the instructional process in microteaching setting by means the factor analytical r-technique: (a) to identify the structural dimensions of interaction; (b) to consider whether they correspond to the logical dimensions of the theoretical framework; (c) to consider the behavior of the emerging factors (factor scores) in combination with certain other variables, frame factors as classified in accordance with the sequence of microlessons; (d) to describe the baseline interaction and its development trends within a revised microteaching course comparing the lesson groups. (e) In the second phase investigate by using replicated designs the stability of the factor structure.

12.3.3 Methods

Design and material

In the first phase of this validation study, the data were 221 microlessons recorded during the revised microteaching course (n=74) and grouped by the sequence of subject's microlessons and observed from video recorded material (Appendix 6.1) by using PEIAC/LH II category system (Table 37, p. 199). A microlesson has been used as the statistical unit and the multivariate method used allowed the treatment of class x period. The primary data consisted of the double codings of six-second intervals in each of the 16 categories in each lesson. To compensate for the variance in the time of microlessons, percentages were used as the basis for the correlations. Reliability (.78) was estimated by means of Scott's Pi coefficient (see Table 39, p. 200).

In the second phase of the validation study replication of the earlier design was used. Data were 126 microlessons recorded during the 1988 microteaching course (n=42) (Appendix 6) and observed from video recorded material 125 microlessons, 10500 6 sec time units by outside trained observer. The reliability of coding was estimated on the basis of 13 cases 1300 6 sec time units; Inter-coder agreement cluster I .76, II .84; within-coder constancy I .98, II .98 and between-coder constancy I .73, II .98) - (Pi .79). The results in item level and the

correlation matrix and factor structure is presented in Appendix 6. The correlation matrix was examined using the determinant coefficient (.0013). In order to determine which of the three lesson groups of the male, female and total population differed from one another, multiple range tests, Schéffe, $p < .05$ were conducted: T-values were calculated after applying Barlett's (1937) test for homogeneity of variance. (Heinilä 1988)

12.3.4 Results of factor analysis

Correlations

The correlation matrices between categories of the two clusters (Table 43) express the relative independence of the categories of both clusters throughout the lessons.

TABLE 43 Pearson correlation coefficient between the PEIAC/LH-75 II categories across three-microlesson observation (n=221).

Cluster category. No																	
I	01																
	02	30															
	31	21	03														
	32	-01	-16	47													
	33	04	-16	35	07												
	41	-29	-18	-28	-16	-29											
	42	13	-17	45	39	34	-29										
	51	-25	-20	-71	-41	-37	33	-45									
	52	-11	-04	-34	-30	-25	12	-37	17								
	06	-18	-06	-39	-19	-23	41	-42	37	05							
	07	06	21	-13	-16	-12	-09	-19	07	18	02						
	08	-01	27	26	-02	16	11	01	-30	-10	-05	10					
	09	07	24	-01	-04	-09	-09	-17	01	05	08	25	-07				
	10	-08	-20	25	26	28	-50	42	-58	-23	-62	-21	-13	-26			
	11	-04	06	-02	-05	-03	02	-07	-04	-01	-02	00	07	-05	06		
	12	-08	-08	-15	-06	-06	-03	01	05	20	07	-02	-05	-03	-08	-01	
II	1	-26	-22	04	-05	10	41	-03	09	07	08	-10	19	-06	-23	-01	05
	2	26	22	-04	05	-10	-41	03	-09	-07	-08	-10	-19	-06	23	01	-05	-1.00
		01	02	31	32	33	41	42	51	52	06	07	08	09	10	11	12	12
		I														II		

However, it can be noted, that one category, 4.1, (teacher asks question, initiates, terminates activity) correlated highly (.41) positively and negatively with the two categories of the second cluster (pupil collective passivity/activity). This is quite natural, because it has two meanings and it forms a 'bridge' between talk and movement. Also, using ipsative nominal scales, it is evident that there will be some high negative correlations: because the process is always in some state, an increase in any one form of behaviour

leads necessarily to a decrease in the other forms. In the second cluster it was evident and the two categories correlated negatively. Also, in the verbal cluster (I) the category indicating teacher's silent behaviour (I/10) and the category indicated the most dominant teacher's verbal behaviour (I/51) correlated negatively. It is also understandable that there will be negative correlation between the categories of teacher initiative behaviours and response behaviours. The highest positive correlation (.47) was found between subscripted categories 3.1 and 3.2, and also positive high correlation (.45) between subscripted categories 3.1 and 4.2. The correlation between 4.2 and 3.1 was quite logical. Psychologically, however, they may by thought to be near each other's. In generally the figures are so low that categories may be considered sufficiently independent of each other to meet the requirements of independence imposed on observational methods.

Factor structure

Correlation matrices were factored by using the principal axis method, and the numerically highest correlations were used as estimates of h^2. Rotation was carried out by the varimax technique. This rotation method was chosen because, being orthogonal, it was likely to yield a simple and clear cut result useful at the initial stage of this "structure seeking" longitudinal investigation.

Three factors proved to be the most interpretable and stable combination. The factors extracted were interpreted as a structural dimension analysing them, and naming the composite pattern.

Factor analysis yielded three factors accounting for 39% of the total variance (Table 44).

From the point of view of category-construct correspondence it was interesting to study how loadings of some categories were spread over the dimensions, i.e. how much or little the categories fill the demand of homogeneity and unidimensionality. Only three categories seemed to have cross loadings. Category 4.1 (teacher asks questions, initiate terminate activity requiring narrow answers - initiates shot-term activity, terminates activity) - is clearly loaded on two factors. It might be best to handle it as two different categories. It might be too narrow. Also category 10 - teacher follows pupils activity, silent guidance – had a negative loading in one factor, and low negative loadings in two other factors. This category dealing with actor and the channel of communication, as well as pupil's verbal response behavior category 08, is best to handle as two different categories. All structural dimensions needed several elements, categories, to be constructed.

It can be concluded that most of the categories filled the demand of unidimensionality and homogeneity, when the interpretation of factors is kept in mind.

TABLE 44 Factor analysis of students' process behavior (PEIAC/LH-75II) variables accross three successive microlessons (n=221; n=74)

Cluster categories No		Factor loadings 1.	2.	3.	h^2
Teachers initiations (+) vs response behavior (-)					
I					
5.1	Presents information, opinions, demonstrates movement patterns, makes a pupil demonstrate	.80	.02	-.25	.70
3.1	Makes use of the ideas and movement patterns suggested by a pupil: clarifies, expands, builds questions and movement initiations on the ideas expressed by a pupil	-.79	.14	.21	.68
4.2	Makes questions requiring higher level of thinking and activity	-.65	.06	-.20	.46
10.	Teacher follows pupil's activity, silent guidance	-.63	-.33	-.37	.64
11.	Teacher silent participation in movement activity	.00	-.01	.03	.00
6.	Gives directions, commands during activity (pupils expected to comply)	.59	.17	.11	.39
3.2	Summarizes pupil's ideas and movement patterns	-.50	-.03	-.13	.27
3.3	Compaires the ideas or movement patterns expressed by one pupil to those of another or to those given, repeats pupil's ideas, asks a pupil to demonstrate	-.49	.10	-.08	.25
5.2	Organizes pupils, material division of labour and responsibility	.40	.02	.04	.16
Channel of teacher-pupil communication verbal(-) vs. - motor (+)					
II					
2	Pupil's collective movement activity/passivity	-.02	-.97	-.18	.97
4.1	Teacher asks questions requiring narrow answers', initiates and terminates short-term activity	.45	.46	-.13	.43
I					
Teacher silence vs. – motivational communication (+)					
2.	Gives corrective feedback, directs, clarifies answers pupil's questions	.03	-.12	.62	.40
1.	Accepts, praises, encourages	-.18	-.12	.39	.23
7.	Criticises pupils behaviour, rejects movement	.19	-.06	.37	.18
9.	Pupil initiates speech, asks for instructions, express own ideas or movement patterns	.12	.07	.37	.15
8.	Pupil answers questions/made by teacher	-.19	.31	.35	.25
12.	Confused situation, uproar	.12	.01	-.12	.03
Eigenvalue		3.4	2.4	1.4	7.2
% common variance		46.4	34.4	19.2	100
% total variance		18.1	13.4	7.5	39.0

The first Factor obtained explained 46.4% of the common variance and was bipolar, clear cut in content as shown in the Table 44. Nine categories of the first cluster loaded heavily on Factor 1. The negative pole concerned the teacher's verbal nondirect communication (3.1 -.79; 32. -.50; 33. -.49; 4.1 -.65) and silent guidance (-.63), whereas the positive pole was associated with teacher's direct behavior as presenting information (.80), organizing pupils (.40), gives direction

commends during activity (.59). It was labelled *"Teacher initiation (+) vs. - teacher response behavior (-)"*.

Factor 2 loaded heavily on two categories from the two clusters. The dominating characteristics of this constructional dimension were the form of pupils' behavior: movement passivity/activity (.97) and teachers initiation and termination of the short-term activity (.46). In the negative pole the highest loading was pupils collective activity (.97) and it was related to teachers silent guidance (-.33). It was labelled *"channel of teacher-pupil communication: verbal (-) vs. motor (+)"*.

In Factor 3, all the highest loadings were positive. The dominant characteristic was teacher-pupil verbal communication. The dimension was typified by the high loading, of teacher's corrective feedback (.62), acceptance (.39), and pupil's verbal initiations (.37) and responses (.35) and also by teacher's criticism (.37). The negative pole was typified by loading of teacher silent guidance (.37). It was labelled *"Teacher silence (-) vs. Teacher feedback and motivational communication (+)"*.

12.3.5 The variance of factor scores by microlesson groups

Differences between microlessons were studied in relation to factor structure: the behavior of resultant factors was considered in combination with frame factors as classified according to the order of microlessons.

Comparison between results obtained in two data sets (n=221, n=148) was conducted. In the second factor analysis, the order of factor 2 and 3 was changed (Appendix 6). The summarized results are presented in Table 45 and in Figures 30 and 31.

TABLE 45 The comparison of the three microlessons factor scores among the students of the two course groups 1 (n=74) and 2 (n=42) ANOVAs and Schéffe multiple Range test

Factors			Microlessons 1st	2nd	3rd	ANOVA df=146 1-2 t	df=146 1-3 t	df=14 2-3 t	df=219 F	Schéffe *) test,p< .05 df=126 t
1. Teacher initiation (-) vs. response behavior (+)	(1)	M	-.40	.54	.55	10.8***	11.4***	1.23	82.15***	1-2
		SD	.74	.73	.81					1-3
	(2)	M	-.86	.32	.54					1-2
		SD	.79	.85	.64					1-3
2. Channel of teacher - pupil communication verbal (-) vs. motor (+)	(1)	M	.50	.50	.50	0.20	0.36	0.17	0.93	
		SD	.96	.96	1.09					
	(2)	M	.14	-.21	.08					
		SD	.69	.85	.78					
3. Teacher silence (-) vs. teacher feedback and motivational communication (+)	(1)	M	.46	.52	.53	3.79**	4.34**	0.81	11.38***	1-2
		SD	.96	.89	1.02					1-3
	(2)	M	-.41	.13	.28					1-2
		SD	.75	.89	.90					1-3

** = p< .01, *** = p< .001

*) t-values were calculated after applying Barlett's (1937) test for homogeneity of variance in the second dataset.

Generally the internal consistencies of factors were good and variance of scores in original scale was F 1.90, F 2.98 and F 3.98. The intercorrelations of varimax factor scores estimated by lesson groups were low (.00, .02, .07).

Comparing variance of factors 1 and 3 ANOVAs among lesson groups showed significant differences ($F = 82.15$, $df = 219$, $p < .001$ and $F = 11.38$, $df = 219$, $p < .01$), between the first (control) and the other two microlessons, but not between the two last microlessons. The scores of the second factor engaged to MET, did not differ significantly between microlesson groups as expected. The comparison of the variance among factor 1 and factor 3 scores indicated that teacher response behavior as well as the motivational communication and feedback increased significantly in the two least microlessons. In the second analyses the variance of factor 1 ($F=3.79$, df 123, $p < .001$) and Factor 3 ($F= 4.34$, df 123, $p < .01$) scores was analogous.

The fact that the lesson groups could be placed dimensionally (direct - nondirect teaching) and contextually (non-verbal - verbal communication) is interesting from the theoretical point of view and the cognitive orientation of the study (Flanders, 1965, 1970, Heinilä, 1977b). Influence students teaching behavior could be verified. Results of the replicated pilot study (Heinilä 1988) supported these findings; the consistency of factor structure in the two different data set was stable, especially, and the explanatory power of factor 1 "Teacher Initiation (+) vs. response behavior was high, 46-52 percent of the common variance explained, and its sensitivity to determine the variance between lesson groups was good.

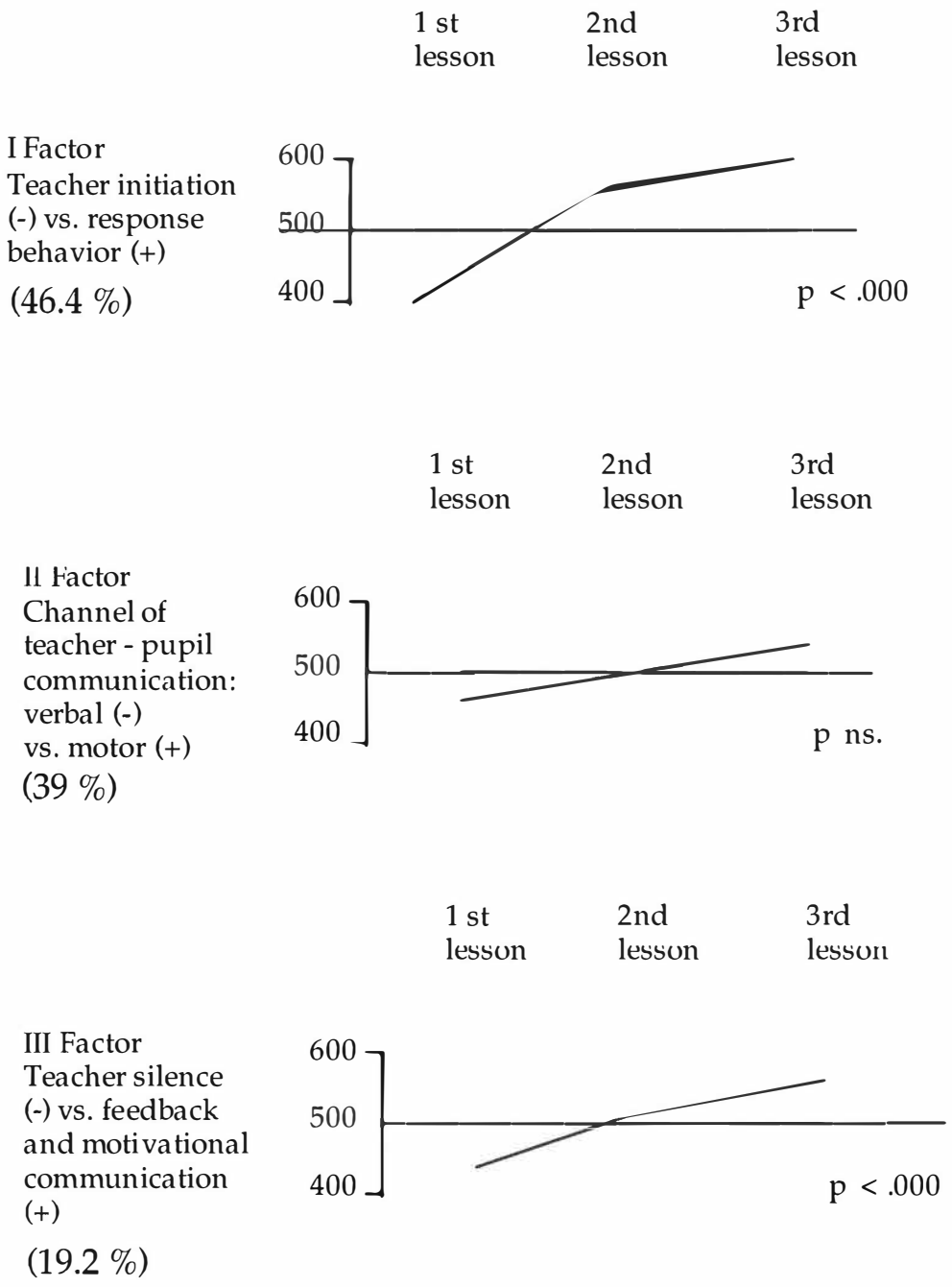


FIGURE 30 Location of microlessons (n=221) in structural dimensions based on the means and dispersion of factor scores

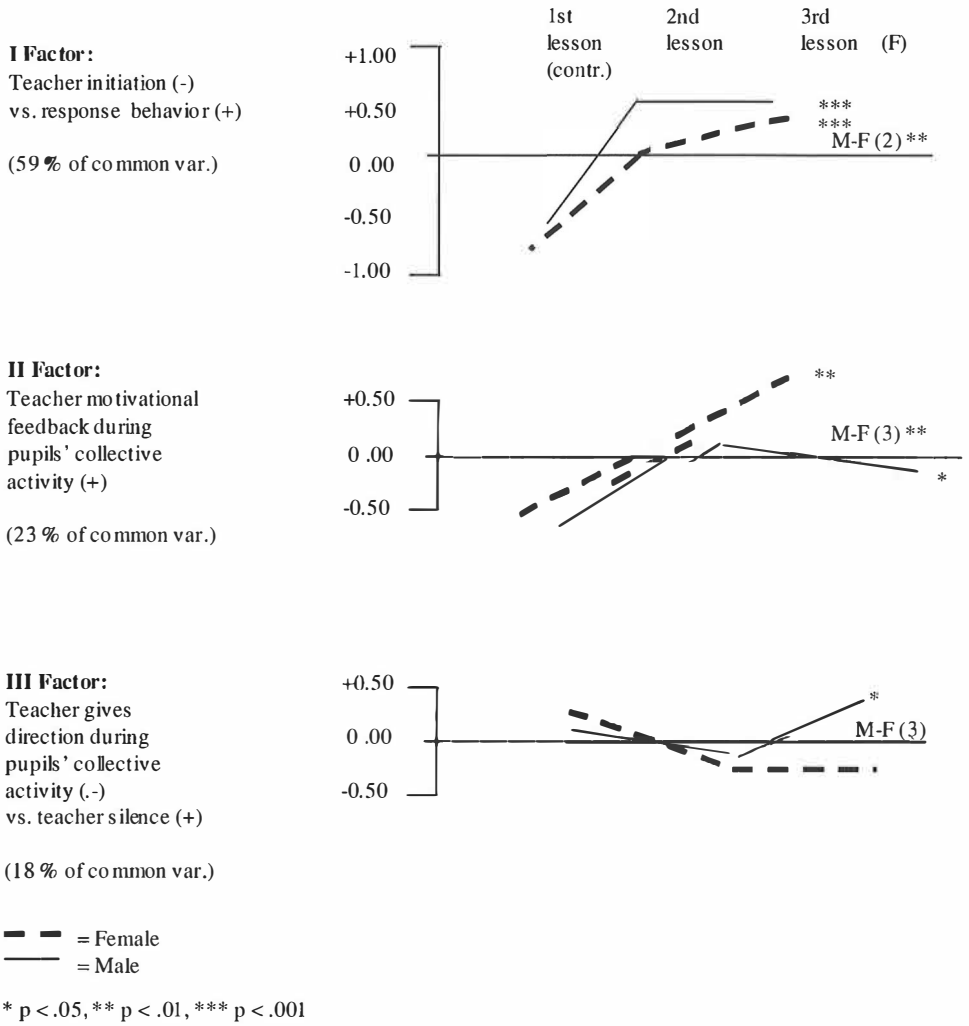


FIGURE 31 Average location of male (n=21) and female (n=21) students' three microlessons in structural dimensions based on the means and dispersion of factor scores (n=126)

As shown in Figure 31, comparing the means of the male and female students microlessons factor scores, there were found statistically significant differences in learning gain: e.g. the male students' behavior changed more quickly in the first structural main dimension; and in the two other, there were significant differences in the third lesson between gender groups: the female students were giving direction and feedback during acitivity more than the male students.

12.3.6 The intercorrelations between PEIAC/LH-75 II variables, indices and factor I scores

As shown in Table 46 it was also found that Pearson's two-tailed correlation coefficients between the ID-index (teacher's response behavior) and Factor 1 (teacher initiation vs. response behavior) scores were high (r .74) and statistically significant at one percent level. The correlation of ID index between microlessons 1 (control) and 2-3 was low (.06) whereas between the 2nd and 3rd high, (r .93 and r.87) and statistically significant at 0.1 percent level. Also the correlation between these PEIAC/LH-75 II variables for the male was high (r .83) and statistically significant at one percent level, whereas for the female lower (r .48) and statistically significant at five percent level. Figure 32 presents comparison of selected indices (PEIAC/LH-75 II) of the second data sets observations (N = 42, 126 microlessons).

TABLE 46 Two-tailed inter-correlations between ID-indices of 1st, 2nd, 3rd microlessons and sum scores of 2nd and 3rd microlessons and between F1 scores (criterion variables)

Variables No	(1)	(2)	(3)	(4)	F1-scores all	M	F
(1) ID-index, lesson 1	-						
(2) ID-index, lesson 2	.11	-					
(3) ID-index, lesson 3	-.02	.63***	-				
(4) ID-index,, lesson 2 and 3	.06	.93***	.87***	-	.74**	.83**	.48*

*p<.05, **p<.01, p<.001
M=male
F=female

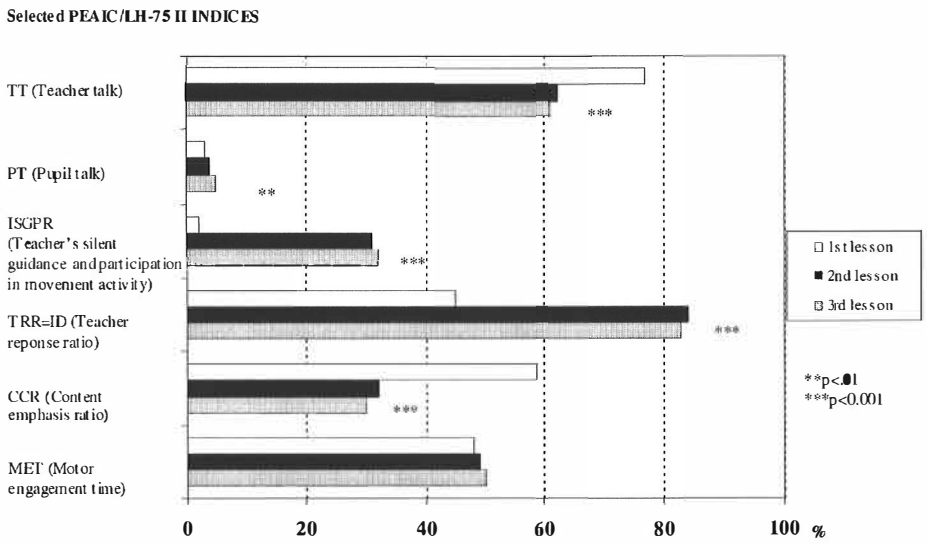


FIGURE 32 Comparison of lessons 1 (control), 2 and 3 percentage index means

12.3.7 Discussion and conclusions of Phase II, B

It might be concluded, based on results of factor analysis by using r-technique for analysing the 221 and 146 video recorded, coded (reliability, Scott's Pi .78) microlessondata sets, that the structure of factors obtained was clear and it explained 39%-34% of the variance; the latter without cluster II. Also the factor structure was stable. The modified observation instrument PEIAC/LH-75 II was also sensitive to discriminant aspects of directive-non-directive teaching behavior of lesson groups (1., 2. and 3.) as assumed as well as aspects of non-verbal and verbal communication (cognitive, affective and psychomotor proprieties of the content). Factor 1,"Teacher initiation vs. response behavior", explained the main part, 46- 52 percent, of the common variance between the two data sets. Lesson groups could be placed in a certain location in on the dimension which reflected their aspects of direct-nondirect teaching in a context of verbal-nonverbal and movement communication. The structural construct of Factors is interesting from the point of view of the theory (Flanders, 1965, 18, 1970) and the PEIAC/LH-75-System presented in Framework, Paradigm: Activity Forms in the Paradigm of PEIAC/LH-75, Figure 20 (p. 170). Also the aspect of pupil's motor engagement time (MET) was integrated to the structural frame, as assumed. The congruence between objectives and degree of their realization could also be assessed.

Factor 1 was found to be composed of the same variables as the Index TRR (=ID-index) teacher's response behavior ratio (Table 43 and 44): the correlation between F1 scores and ID-index was high r .74 and statistically significant at one percent level. The ID-index was used as criterion variable also in studies by using Flanders (1970) Interaction Analysis System (FIAC).

Also based on the results obtained in pilot studies conducted in natural setting by Reponen (1979), it was established that (1) the order of PEIAC/LH-75 (Heinilä, 1977a, 1977b) indices revealed differences between experienced teachers (N = 2, 24 lessons) with regard to the rank order of behaviors and (2) between two group of student teachers (n=44, 14 male and 30 female), and between student teachers and experienced teachers. These results support the results of this validation study.

It was concluded that (1) the 18-category PEIAC/LH-75 II, system for combining process and cognitive orientation possesses a definite degree of construct validity, and that (2) it is sufficiently sensitive to discriminant aspects of direct/nondirect teaching behavior in a microteaching setting and aspects of direct - indirect teaching behavior and (3) discriminant also the aspect of pupil's motor engagement time (MET).

12.3.8 Summary and conclusions (Phase II A and B)

In the second phase of program evaluation in pilot studies A and B, the goal was to assign the basic elements within the components of the multiple baseline curriculum evaluation design, (modified measuring instrument PEIAC/LH-75 II and criterion tasks, the six non-directive teaching models). The PEIAC/LH-75 II taxonomy contains two clusters (teacher's speech, silent and movement and

pupil's speech (16 categories), 2) pupil's collective movement activity/passivity (2 categories). Double coding was made at six-second intervals by trained outside observers (Scott's Pi .78).

In the first pilot study the data consisted of 148 microlessons observations grouped by teaching models and in the second pilot study of the observations of 221 microlessons 1, 2 and 3 (n=74) and in the replicated design of 126 microlesson observations (n=42). The construct validity and sensitivity of the measuring instrument was estimated by using a multivariate approaches discriminant analysis and factor analysis.

Based on the results of the two pilot studies (A and B) here it was concluded that the 18-category PEIAC/LH-75 II-system for combining process and cognitive orientation used in interventions in connection of the revised microteaching course, proved feasible both for research and for teacher training: (1) It possessed a definite degree of construct validity, and (2) it was sufficiently sensitive to discriminant aspect of direct/nondirect teaching behavior in microteaching setting in Physical education and (3) it included a definite degree of aspects of nondirect teaching behavior operationalized as teaching models. It took into the account clear specification of the task to be learned as presented in program intervention strategy design (Figure 27, p. 213 -214) and measured students' engagement with the criterion task. It facilitated the operationalization of intended behavior, code patterns, and helped to teach discrimination and create desirable teaching patterns as defined in connection microteaching course based on the theoretical Framework (Flanders 1970, Heinilä 1977a, 1977b, 1990). Also based on results of factor analysis and of comparison of factor 1, "Teacher initiation (-) vs. response behavior (+)" scores and ID-index correlations (r .74, $p<0.01$) the claim of the construct validity of PEIAC/LH-75 II system was supported. Furthermore, in comparing the results obtained in investigations conducted with different data sets and replicated design, the stability of factor structure was assessed and it supported the assertion of the construct validity of PEIAC/LH-75 II used in microteaching course (Heinilä 1988).

Also the results of explorative investigations conducted in natural settings by using the PEIAC/LH-75 I and II system by Akkanen (1979) and Reponen (1979), supported findings of these validation studies.

Although the results of these multivariate analyses still can only be regarded as tentative due to the nature of the level of the measurement scale, they yielded quite useful information for the further development of the revised course program as well as the instrument used in interventions. The multivariate analysis conducted helped for refining the discriminative model and reducing data, variables to be used in a long-term multidimensional research project in program predictive validation.

12.4 Pilot study II C: The teaching behavior rating scale: reliability and validity

12.4.1 Introduction

(1) Background and purpose

Within the framework of this on reliability and validity, it was important to device an instrument, which would provide measure of the students "entry teaching behavior". The term used here, refers to specific, explicit, observable and also countable behavior in educational settings. Entry skills and knowledge are those behaviors that students must be able to perform before receiving new instruction. These behaviors are usually determined through the use of instructional analysis techniques such as category systems or rating scales. Gage (1969) and Rosenshine (1971) called the types of observations that might be obtained "low inference variables" – usually tabulation of observed teaching behavior - and "high inference variables" – usually perceptions of obtained on rating scales. The classification is based on the amount of inference required of observer or person reading the research report. (Dunkin & Biddle 1974, Rosenshine 1971.) For the comprehensive understanding of classroom social interaction and acquisition of teaching skills, it is recommended to study the both variables simultaneously and to discover the linkages between the two (Rosenshine 1971). This was the research strategy used in this study.

The teaching behavior rating scale was aimed to be used as a means for determining the students' entry level teaching skills, firstly in students intake test (microteaching episode, 3.5 min) and secondly two years later before the course of Microteaching, (microteaching episode, control 5 min) and also as a form of intervention in the course of microteaching. The developed rating scale was used in the Faculty of Sport and Health Sciences from the year 1976 onward starting simultaneously with the course of didactic observation and microteaching (1974).

The main objective of this particular study was to determine the reliability, validity and sensitivity of the teaching behavior rating scale.

(2) The teaching behavior rating scale and its development

The Rating Scale was constructed in response to the Faculty board members' interest in developing a better oral delivery of students and to some degree, interest to implement students' process behavior – interaction-related to various features of the teaching situation such as the tasks, the size of group, gender and age of participants, and overall the effectiveness of teaching (Heinilä 1988). Furthermore, the rationale for the construction of the rating scale was based on the author's observations of physical education teaching (Heinilä 1970) and on observations on students' oral delivery tests and on students' pre-service teaching practice where failures in presentation, flexibility in process behavior and in creativity were evident. The rating dimensions selected to the measuring

instrument were based on research results, obtained in the study on the relationships between teachers personality and teaching behavior and also on effectiveness of teaching (Flanders 1965, 1970, Hytönen 1973, Hytönen & Komulainen 1973, Kane 1968, Medley 1971, 1987, Rogers 1967, Rosenshine 1970, 1976, Rosenshine & Furst 1971, Whitehead 1980) and also on studies on acquisition of different teaching skills (Siedentop 1981).

Furthermore, it must be noted that the Curriculum of the Faculty was based on "Human Interaction Model" (Heinilä 1988, Telama 1979) and, therefore, both the intake procedure tests and preservice teacher education programs and their components were assumed to be congruent with it.

The teaching behavior rating scale contains four items, six-point scale. Information on it and its use in rating and in connection with teaching episodes is presented in Appendix 7. The instructions for teaching episodes were given for subjects 10 minutes before for preparation.

The rating dimensions selected were connected with following behavioral characteristics: the first item, on teachers' "*Presentation*: voice quality, expression, fluency, clarity, movement behavior"; the second item, on content presentation: "*Understanding of task content, phases and instruction*"; the third item on "*Communication: interaction with pupils,* directiveness of main points, observation, feedback"; and the fourth item on "*Creativity:* originality, aptness, presentation of main points".

12.4.2 Research task

The first research task was the selection and development of measuring instrument for diagnosing subject's entry teaching behavior; and the second to answer the research question: does the measuring instrument have some conceptual integrity and coherence with frame of the course program and student intake procedure.

The three subquestions were:

1) Assessment of the reliability and validity of the measuring instrument.
2) What is the stability of the measuring instrument and the scale scores and what differences exists within and between subscale scores variance among the male and female subpopulation groups?
3) What is the usefulness of those variables for research and for further program evaluations, its internal and external predictive validation?

12.4.3 Method: Procedures and instrumentation

Data: the reliability and validity of the rating scale was assessed by using video recorded material of the microteaching episode observations (5 min x 75 and 5 min x 42) and summed-scores of the subject's microteaching episode (3.5 min) ratings in students' intake taken from the documents of the Faculty. The raters and observers were trained, outside observers, post-graduate university lecturers.

The material, total percentage of frequencies summed per each item over the sample of teaching episode was ranked separately by item scale (1-6).

The statistical processing of the material was conducted with subprogram Fortran NMCC and SPSSx at University of Jyväskylä, Statistical Institute.

Reliability of ratings, inter-rather agreement and stability (Siegel 1956) was determined by means of Kendall's Coefficient of Congordance, W and rank order correlations coefficients, r or and for Chi Square test estimating the statistical significance of coefficients obtained.

Stability of ratings was determined by comparison of results of repeated tests, between four weeks and two years interval based on W values rank order correlations, r, Chi Square tests and also on Pearson's Coefficients of correlation and two-tailed t-test. Validation of the rating scale was based on results obtained by using two measuring instruments for determination of students' entry teaching skills in teaching episode (control) from the video recorded material (n=42). Rathers and observers using PEIAC/LH-75 II category system were different persons, trained outside observers. Reliability of PEIAC/LH-75 II was assessed by means of Scott's pi coefficient (.78). The estimate for validity was Pearson's Coefficient of correlation (r) and two-tailed t-test used in comparison of the level and statistical significance of the relationships between rating scale scores, item 3 (teacher-pupil interaction) and ID-index (teachers response behavior) scores and also item 3 and F1 ("Teacher's initiative vs. response behavior scores") variables validated in the earlier pilot study II b. The homogeneity of the variance was defined by means of Barlett's test (1937) t-tests and by comparison of the variance across eight subpopulations (n=205) conducted by using Schéffe's Multiple Range test (1959) and ANOVAs, one-tailed t-tests (Appendix 7.3).

12.4.4 Results

Reliability of the rating scale (1)

Based on results presented in Appendix 7.2, the reliability, inter-rather agreement was in the first test (n=75) rather high: the median value of summed items was W .75 and the rank order correlations between raters (r^2) were statistically significant at 0.01 percent level. In the second investigation (n=42), the median value of reliability index was .68 and the rank order correlations between raters, r^2 = .36, statistically significant at the one percent level. On the item level, the indices ranged in the first test between W .73 - .75, (Chi Square .46-.51, df = 74, p< .001) for four items and in the second test between W .56 - .70. The rank order correlation coefficients ranked from r^2 = .46 to r^2 = .57 and they were statistically significant or beyout (Chi Square, df = 41, p< .04; p< .28; p< .05, p< .09).

The rank order correlation between the two raters' total summed ratings was r^2 = .36 and the Pearson's correlation coefficient was r = .53 and statistically significant at one percent level.

Stability of the rating scale (2)

The stability of ratings was determined firstly, based as results obtained of the second observations four weeks later (n = 10, n = 12). In the first test reliability coefficient for summed scores, Median value was W .84, r^2 .56 (Chi Square, df =9, p< .05).

Secondly, the stability of ratings as shown in table 47 was statistically significant at one percent level (r = .52). It was based on correlations between two rathers summed scores gathered of the same population (n=42) between two years interval, and determined by means of Pearson's Coefficient of correlation and t-tailed t-test. Furthermore, item level the correlations between the two teaching episode test summed scores ranged between, r = .56 – r = .39 and were statistically significant at one percent level.

TABLE 47 Two-tailed correlation coefficients between rating scale scores gathered at video recorded episodes (control) 5 min x 42 and two years earlier in student intake (sum scores) and between PEIAC/LH-75 II observation system variables, ID-index and F1-scores (n=42)

Items	Intake	ID-index	F1-scores
1. Presentation: - voice quality - expression fuency - clarity of movement behavior	.40**	-.14	.31*
1. Understanding of task content - phases in instruction	.56***	.15	.12
2. Communication - interaction with pupils, directiveness of main points - observation - feedback	.39**	.42**	.30
3. Creativity - originality, aptness, presentation of main points	.40**	.05	.04
Sum scores	.52**	.18	.11

* p < .05, ** p < .01, *** p < .001

And thirdly, the stability of ratings was verified among the total sample (n = 205) (see Appendix 7.4).

Validity of the rating scale (3)

The validity of the rating scale was determined by comparing the results obtained with two measuring instruments used as the means for assessment students (n = 42) entry teaching skills in a microteaching episode (control 5 min). As shown in Table 47, the correlation between the scores of the estimate of validity, rating scale item three (teacher-pupil interaction) and PEIAC/LH-75 II category system variable, ID-index (teachers' response behavior) was r = .42, statistically significant at one percent level; and for the male (n = 21) r = .44 and female (n = 21) r = .42, both statistically significant at the one percent level. Factor F1 ("teacher initiation vs. response behavior") scores correlation between item three scores was r = .30 statistically nearly by significant and between sum

scores r = .36 statistically significant at one percent level and also the correlation between ID-index and F1 scores was high (r = .65) and statistically significant at one percent level (Appendix 7.3).

The sensitivity of the rating scale (4)

In the final stage of the study the homogeneity of the variance was assessed for the ratings by means of Barlett's test (1936); and the comparison of the variance between different populations was made by using Schéffe (1959) Multiple Range test and by means of two-tailed t-test across subpopulations: the four didactic observation and microteaching intake-course groups (1976, 1979, 1980, 1986, 1988), and for the male and female populations and also across gender course groups grouped into decade 1970's and 1980's groups. As shown in Table 48, e.g. the results concerning the male population scores showed that there were statistically significant differences in all items and also in sum scores (at one percent level). The male entry teaching behavior level was significantly higher in 1980's than in 1970's. Also the total sample scores differed in the same direction, e.g. in the first item "Presentation", statistically significantly difference at five percent level of confidence was identified.

TABLE 48 Description of subpopulations: comparison of students entry teaching behavior ratings scale frequencies (means and standard deviations) in the microteaching episode (control 5 min x 205) between combined group 1. and 2. (1970's and 1980's) among male, female, and total population, two-tailed t-tests

Variables	Male			Female			Total		
	n=47 1970	n=37 1980	t	n=75 1970	n=46 1980	t	n=122 1970	n=83 1980	t
1. Presentation	3.7 (0.9)	4.3 (1.0)	**	3.9 (3.9)	4.0 (4.8)		3.8 (0.9)	4.1 (0.9)	*
2. Understanding of task content	3.7 (0.9)	4.1 (1.0)	*	3.9 (0.9)	3.8 (0.8)		3.8 (0.9)	3.9 (1.0)	
3. Communication teacher – pupil interaction	3.8 (0.9)	4.2 (0.3)	*	3.8 (1.0)	3.6 (1.2)		3.8 (1.0)	3.9 (1.0)	
4. Creativity	3.1 (1.1)	3.4 (1.3)		3.2 (1.2)	3.1 (1.3)		3.2 (1.4)	3.2 (1.3)	
Sum scores	14.2 (3.0)	16.0 (3.2)	**	14.8 (3.3)	14.5 (3.6)		14.6 (3.2)	15.2 (3.5)	

*= p<.05; ** = p<.01

12.4.5 Discussion and conclusions – Phase IIC

The teaching behavior rating scale was constructed in response to the Faculty board members' interest in the quality of students' oral delivery and process behavior, with the scale to be used in student intake and two years later in the course of Microteaching as a means of intervention for determining the level of students' entry teaching skills. It contains four items, on a six-point scale. The reliability and validity were determined in studies by using video recorded material (5 min x n=75 and 5 min x n=42) of micro-teaching episodes control before the course of microteaching and also sum scores from the intake test, microteaching episode (3.5 min) observed and rated by two trained outside observers, post-graduate lecturers of the University.

Reliability, inter-rater agreement determined by means of Kendall's Coefficient of Congordance, W, ranged for sum scores between (MD values) .75 and .68 and were statistically significant (Chi Square) at one and at five percent level in two different tests. On the item level the reliability indices were statistically significant at one percent level (75 cases) and in the second test (42 cases) at five percent level or beyond. *Stability* determined by correlations between sum scores of repeated ratings and between sum scores of student intake, ratings of teaching episode two years earlier was r = .52, statistically significant at one percent level. Also results obtained in the other samples ranged between r .30 – r .23 and were significant at five percent level.

Validity of the ratings scale was determined by compairing results obtained from the same material, the first microlesson (control) observations with two instruments, the rating scale variables item 3 ("teacher-pupil interaction") and PEIAC/LH-75 variables, ID-index (teachers' response behavior) and Factor 1 scores (teacher initiation vs. response behavior). The Pearson correlations were r = .42 and r = .30 and statistically significant at the one and at five percent level. The intake teaching episode test sum scores correlated with the ID-index scores at r = .34 and with F1 scores r = .36, statistically significant at five percent level.

Moreover, defining the homogeneity of the variance of the scores in eight different subpopulations (n=205) and by comparing the variances between population groups it was concluded that the rating scale was also sensitive for assessing the level of students' entry teaching skills. Also the congruence between the level of the objectives and the level of observation of the test results in two contextual settings was verified. *The goals The Faculty Teacher Education program and the course were congruent concerning e.g. implementation of student's oral delivery.*

These findings paralled with the results obtained in the earlier studies of this project (Heinilä 1988). Moreover, e.g. of Rosenshine and Furst (1971, 44) (who interviewed results of teaching effectiveness) recognized that in seven experimental studies where students or observers ratings were used as the criterion measure the correlations of "clarity of presentation" were statistically significant and ranged between .37 - .76.

Based on results obtained it was deemed that the reliability and validity of the rating scale was sufficiently high for further analysis and to be used means of student intake and interventions in the course of Microteaching.

However, the faculty student intake procedure where teaching episode was weighted low in total scores, only 11.5 percent (after study degree program reform 1978), and did not correlate significantly with total intake sum scores in most subpopulations of this study (Heinilä 1988) (see Appendix 9.4.1). This suggested that there was a need to consider what this might mean for further program evaluation and for criteria based on program predictive validation.

12.5 Phase II. Pilot study D: Students' attitudes, "ideal" P.E. teacher expectation rating scale reliability and construct validity. A multivariate approach

12.5.1 Introduction

Background and need for the study

Within the framework of this particular study with the aim the course program predictive validation it was deemed important to assess the construct validity and sensitivity for the program of the rating scale students' entry attitudes. The measuring instrument was used as a means of intervention with the aim of enhancing students' intentionality and learning.

As stated before, the course package was based on the theory of Flanders (1965, 1970) and had as its goals knowledge and mastery as well as cognitive understanding of teacher-student interaction, as defined by Heinilä's adaptation of Flanders interaction Analysis System (PEIAC/LH-75, II) (Heinilä 1977b, 1979, 1988, 1990). *It might be noted here that the original FIAC-system is classified in the category "affective" by Simon and Boyer (1970, 371)* (cf. Bloom et al. 1956). The program elements to be evaluated here represent this area. This study unit program as stated in introduction of Section II, the course of didactic observation and microteaching, was conducted at the Faculty of Physical and Health Education of the University of Jyväskylä in 1974 – 1991 and in two different kind of contextual settings before and after the study degree program reform (1978).

This study arose out of an assumption, based on observation and research reports, that teachers' personality can have an important effect on both the process and product of teaching (e.g. Hytönen 1973). Kane (1968) also showed that there were no significant changes in personality across three years in teacher training. The "ideal" coach stereotype have been studied and related to measures of personality and subjective self-assessment of the coaches (Hendry 1969). Moreover, studies connected to teacher education programs aimed at the acquisition of teaching skills have shown that the subjects' characteristics strongly influenced the effects of experimental treatment (Hanke 1980b,

Siedentop 1981) and it has been found that subjects involved in these kind of experiments were characterized with different sensibilities in comparison with the objectives and contents of P.E. teacher education programs. (Siedentop 1986, Silvennoinen et al. 1991, Telama 1970, Telama et al. 1988). This kind of inter-individual variability underlines the necessity of case studies in which different behavioral modifications and students' acquisition of teaching skills can be assessed in terms of the subjects' personal characteristics, such as attitudes and expectations.

Also the subjects' entry narrow teaching style conceptions at the beginning of the microteaching course, as observed by the author, might be an obstacle to success in teacher flexibility training programs. A solution to the problem of promoting students' study motivation and learning may lie in diagnosing subjects' entry attitudes, such as their expectations concerning "ideal" P.E. teacher characteristics, with the aim of widen their conceptions of P.E. teacher characteristics and teaching styles. Because attitudes are learned, they can be changed by education (see e.g. Martin et al. 2001). Awareness of subjects' entry attitudes and their individual, environmental, and time related variation might have potential to help teacher educators and administrators work toward elimination of stereotypes of P.E. teaching behavior – toward flexibility – which was the main purpose also of the didactic observation and microteaching course.

12.5.2 Research task

The first research task was (a) selection and development of the measuring instrument for diagnosing subjects' entry attitudes, their current "ideal" P.E. teacher expectations, and (b) to answer the research question: does the measuring instrument have some conceptual integrity and coherence with the framework of the course program.

The four sub-questions were:

(1) What factors make up microteaching course students' entry attitudes, their "ideal" P.E. teacher expectations, connected to the dimension of student centered- teacher centered teaching style?
(2) What is the stability of the measuring instrument and the scale scores and what differences exist within and between subscales score variance among the male and female subpopulation groups?
(3) What is the sensibility of the measuring instrument to discriminate between student "ideal" P.E. teacher expectations factor scores and contextual variables in terms of students' course group and gender?
(4) What is the usefulness of these variables for research, and for further program evaluation and its predictive validation?

12.5.3 Methods

(1) The "ideal" P.E. teacher expectation questionnaire

The instrument used was an adaptation of the 16 item bipolar, 1-6 rating scale and based on ideas of Flanders (1965) and Rogers (1967). It was developed and validated by Hytönen and Komulainen (1971), to be used in an empirical study conducted in a teaching training course in the subject area of mathematics for controlling the stability of teaching styles studied. The main characteristics are developed based on concepts defined as follows: "Student–centered teaching signifies the kind of teacher behavior in classes where information is transmitted mostly through students' own activity. In teacher-centered teaching the teacher is the active part. A student-centered teacher's relationship to students is characterized by him/her being approachable, as his/her personality gets involved in teaching process. A teacher-centered teacher is distant in relationships with students. An approachable teacher is accessible to all students, uses humour as an instrument in instruction (gets involved) concentrates on activating students. (Hytönen & Komulainen 1971, Rogers 1967, 46, 1980, 114-116).

For the purpose of this study and for diagnosing subjects' entry attitudes before the beginning of training sessions, the questionnaire was deemed to be useful and applicable. However, in the frame of this course on the subject area of physical education teacher training four criterion-based items were added. They were connected to the social form, division of the labour and responsibility and teacher's own motor engagement - variables added also to the PEIAC/LH-75 system. (Heinilä 1977a, 1987, 93-94 and 244)

The "ideal" P.E. teacher expectation questionnaire comparised twenty items in six bipolar response categories is presented in Appendix 8. As stated in information of its use, "the selection of response categories is asking students to imagine the opportunity to choose his/her own teacher based on subjects' "current ideal". Figure 33 illustrates the rating scale and describes results of students' estimations before and after study reform (n = 205).

(2) The statistical procedures

The reliability of the questionnaire was determined in terms of homogeneity, Cronbach's alpha, based on item test correlations for populations among four microteaching course groups (n=205). It ranged in the four-factor solution from .56 to .76 on factor 1, from .45 to .49 on factor 2, from .41 to .56 on factor 3, and from .26 to .52 on factor 4. When population groups were factorised separately, it ranged from .56 to .92 (Heinilä 1988). These figures were judged to be sufficient for further analyses.

For the purpose of this particular study, the material of four replicated case studies (n=205) was subjected to factorization by the principal axis method and to rotation by the orthogonal varimax technique. The comparison of factor variance within and between subpopulation groups was conducted by using one-way ANOVAs and two-tailed t-tests and Pearson's correlation coefficients.

The relations and differences between students' "ideal" P.E. teacher expectation factors and background variables, such as course group and gender, were analyzed by using two-way ANOVAs. The sensibility and power of the instrument was determined by using step-wise discriminant analysis.

The analyses used two-tailed (MD) groups. The assumptions concerning these methods were controlled by using multiple comparison procedures with the following analyses: S-method, (Schéffe –test) and Barlett's test of homogeneity of the variance (1937) (e.g. Dillon and Goldstein 1984); probability level of $p < .05$ was initially set as an indicator of statistical significance in all analyses.

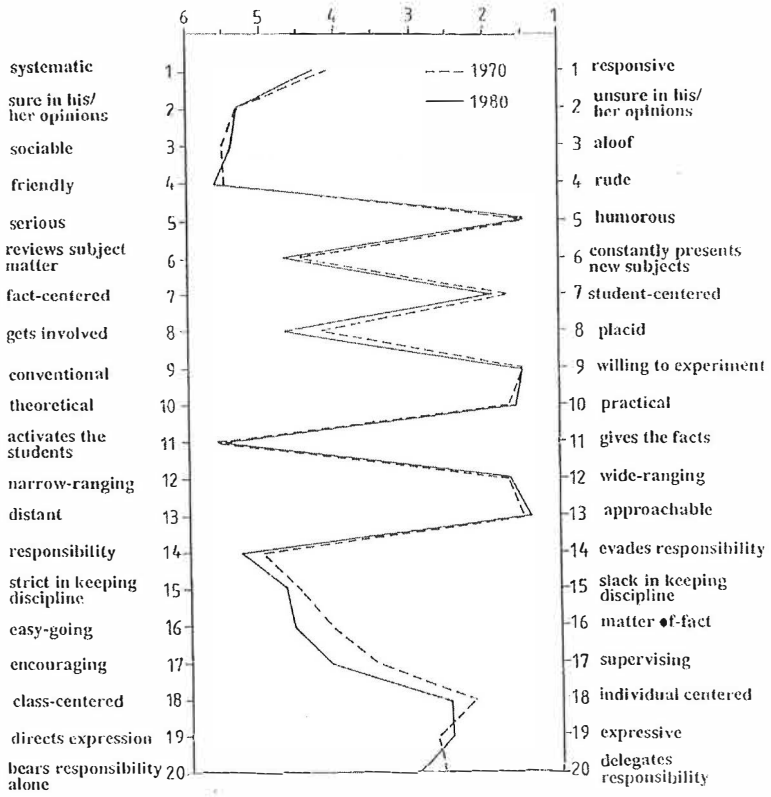


FIGURE 33 Microteaching course groups' estimation of "ideal" P.E. Teacher (n=205) in the 1970s and 1980s

12.5.4 Results

(1) The factorial structure

To have a more efficient rating scale description and to eliminate overlapping of items, the material was subjected to factorisation by the principal axis method, and to rotation by the orthogonal varimax technique. The solution reached with four factors was found to be most clear and interpretable. The

determination of the correlation matrix computed over four course groups (n=205) was .193, which was deemed to be sufficient for further analyses. The criterion used for variables included in factorisation was the minimum communality size of .20, and based on this six items were deleted (2,4,6,10,11 and 20). The four bipolar orthogonal factors extracted only 28.1 percent of the total variance. The structure was similar as in earlier analyses and the variance as rated was higher when factored in each course group separately (32-46%). The correlation between factor scores ranged between (r = 0.2 – 0.9) in the 1980s sample (n=83) (Table 50).

The factors extracted were interpreted as structural dimensions. As shown in Table 49, the first general factor extracted the highest proportion - 14.6 percent - of the total variance and 51.9 percent of the common variance. It was labelled *teacher's personal congruence/genuineness*, in terms of quality that encompasses four traits representing the warmth and "realness" of the teacher (Rogers, 1967). The item with the highest loading (.72) described a teacher using a sense of humour in instruction, as contrasted with being serious. The three other items loaded on this dimension indicated the teacher characteristics of approachable-distant (.35); the second highest was the teacher's willingness to experiment as compared with conventional behavior (.33) and the negative pole of factor dimension loaded item, being sociable and participating in activities in the classroom, compared with teacher's aloofness (-.35).

The second structural bipolar dimension was composed of five characteristics that are general indicators of the social behavior of teacher in the gymnasium. It was called *Social form: class–centered (+) – Individual-centered (-).* The item with the highest loading indicated teacher who spends time encouraging individuals, compared with one who tends to concentrate more on the class as a whole (.61). The second item, which loaded on this factorial dimension, measured the characteristic of direct expression – is self-expressive, participates in activity and spontaneously shares emotions with students (.45). The third item loaded on this factor described a teacher who uses examples from various sources and subjects compared with a teacher who tends to be limited when instructing students and restricts discussion only to the subject-matter compared with a teacher who is wide-ranging (.34). This factor measured a dimension where a teacher is an initiator with students and encourages the whole class purposefully, compared with supervising or simply overseeing their activities (.33).

The third structural factor dimension, called *"teacher involvement (+)"*, was typified by three "ideal" traits: level of teacher's involvement measuring the ability of the teacher to expand into related topics in class, compared with being placid and not digressing from the subject matter (.54); and teachers show a willingness to make decision concerning class when compared with avoidance behavior (.53); the degree to which a teacher is easy-going and related when talking about his or her own experiences with students, in contrast with remaining impersonal in interaction (.45).

The fourth structural dimension, labelled as *"fact-centeredness (+) – student-centeredness (-)"*, describes the teacher's use of authority, as reflected in three aspects: the first item which points out the necessity of teachers to be strict, and

have good disciplinary procedures in a variety of instructional situations as compared with being slack in using management techniques (.41); similarly, being systematic and clear-cut when implementing lessons compared with being responsive in lessons (.41), and an item loaded on the fact-centered – student-centered, paying attention only to the academic achievements of students, when contrasted with behaving in a more student-centered manner (.39).

TABLE 49 Results of principal components analysis on students' attitudes, "ideal" P.E. teacher expectations scale among the four microteaching course (1976, 1979, 1980, 1988), n=205

Items No	Factor loadings				
	1.	2.	3.	4.	h^2
Teacher's personal congruence/genuineness (-)					
5. serious – humorous	.72	.02	-.06	.09	.53
3. sociable – aloof	-.35	-.14	.02	-.28	.22
13. distant – approachable	.35	.28	-.18	-.02	.23
9. conventional – willing to experiment	.33	.33	-.25	.10	.29
Social form: class-centered (+) – individual-centered (-)					
18. class-centered – individual centered	.13	.61	-.05	.01	.39
19. directs expression – expressive	.10	.45	-.09	.06	.23
12. narrow-ranging – wide-ranging	.14	.34	-.32	.16	.27
17. encoraging – supervising	-.11	.33	.22	.19	.21
Teacher involvement (+)					
8. gets involved – placid	.02	-.06	.54	-.12	.31
14. responsibility – evades responsibility	-.07	-.06	.53	.17	.32
16. easy-going – matter of fact	.26	-.06	.45	-.06	.28
Teacher's fact-centeredness (+) – student centeredness (-)					
15. strict in keeping discipline – slack in keeping discipline	-.01	.02	-.10	.41	.18
1. systematic –responsive	-.01	-.05	.09	.40	.17
7. fact-centered – student centered	.30	.25	-.12	.39	.31
Eigenvalue	2.04	0.78	0.58	0.53	3.93
percent of common variance	51.9	19.8	14.9	13.0	100
percent of total variance	14.6	5.6	4.2	3.8	28.1
(determinant of correlation matrix .942928, $p<0.05$)					

TABLE 50 Two-tailed intercorrelation coefficients between students' "ideal" P.E. teacher expectation scales; the 1980s population, n = 83

Scale		(1)	(2)	(3)	(4)
f-1	Teachers personal congruence/ cenuineness	-	.09	.02	.19
f-2	Social form: class-centered (+) – individual-centered (-)		-	.18	-.13
f-3	Teacher involvement (+)			-	.03
f-4	Teacher's fact-centeredness (+) – student centeredness (-)				-

(2) The variance of factor scores among subpopulation groups

The results of one and two-way ANOVAs concerning factor scores is presented in Table 51. By analysing the variance of the four factor scores among subpopulation and gender groups and combined groups representing populations before and after study reform (1978), it can be shown that there were no statistically significant decade course effects in the first general factor. Thus, the students' "ideal" PE teacher personal characteristic expectations *"congruence/ genuineness" issue was consistent over decades and the contextual curriculum and gender variation.* This variable was labelled in an earlier study project "student centered–teacher centered teaching style" (Hytönen & Komulainen 1971). By contrast, in the three other factors, significant contextual effects were noted. In factor three, the variance of factor scores was identified to be related to population group (F, df 1, 14.86, p< .001) at 0.1 percent level. Thus, the "ideal" P.E. teacher of the 1980s, course students, after the study reform, was more "involved" than in 1970's. Also in factor two, a statistically significant decade group effect at the five percent level was found (F = 3.68, df 1, p< .05). The social form used by the "ideal" P.E. teacher was in the 1980s more class-centered, whereas in 1970 individual-centeredness was dominating. In the fourth factor, statistically significant course group and sex effects were found and the interaction between the two was statistically significant (14.42 p< .000). Obviously, the "ideal" P.E. teacher's fact-centeredness was weighted more in the 1980s than in the 1970s, but only in the male population (t = -3.83, p = .001).

TABLE 51 The comparison of students' attitudes, "ideal" PE teacher expectations factor scores among students grouped according to decade (1970s and 1980s) course and gender course groups (two-way ANOVAs, two tailed t-test, n=205

Varimax factor		Decade 1970's (n=122)		Decade 1980's (n=83)			Source of variance:				
		M (n=47)	W (n=75)	M (n=37)	W (n=46)	Sex df=1 F	Decade course df=1 F	Inter-action df=1 F	M-M	T-test W-W	Tot.
F1: Teacher's personal congruence/ genuineness (-)	M	-.04	.03	.08	-.07	0.46	0.00	0.95	-0.73	0.65	0.01
	SD	.81	.83	.76	.70						
F2: Social form: class-centered (+) - individual – centered (-)	M	-.01	-.14	.22	.06	1.83	3.86*	0.03	-1.35	-1.42	-2.05*
	SD	.79	.66	.75	.82						
F3: Teacher involvement (+)	M	-.11	-.20	.16	.30	0.01	14.86***	1.25	-1.60	-3.73***	3.86***
	SD	.72	.75	.79	.66						
F4: Teacher's fact-centeredness (+) - vs. student - centeredness (-)	M	-.36	.07	.81	.01	2.05	6.81**	14.24***	-3.83***	0.54	-2.44*
	SD	.73	.58	.07	.52						

Values are means (SD), F, Fisher's F-statistic (variance ratio); variance are not equal between groups, *, **, *** p<0.05, 0.01 and 0.001 respectively, M=men, W=women

Table 52 reports summarized the results of the variance of factor scores by contextual course group and gender effects, the results of Schéffe test, comparisons of the homogeneity of the factor variance by course group and gender. A comparison of the variance of factor scores between different course

groups showed that the main direction of changes of "ideal" P.E. teacher factor scores was the same as indicated by distributions in an earlier report (Table 51). The "ideal" teacher expectation factors of the course group 1988 differed most clearly from the other groups and especially from the 1976 course group. The statistically significant course group effect on variance in factor scores could be noted in factor two (F = 3.66) at one percent level and in factor three (F = 5.69) at 0.01 percent level. Thus, contextual course group effects were evident, changing from "ideal" PE teacher's individual centeredness to class centeredness and toward teachers' greater involvement. Teacher's responsive involvement is characteristic of the "ideal" PE teacher of the 1988 course and especially of the female student group. Also in these analyses, in factor four, the effect of the two-way interaction gender and course group was evident where teacher's fact-centeredness was weighted more than in other groups, especially among the male students

TABLE 52 The comparison of students' attitudes, "ideal" PE teacher expectations factor scores among the four course and gender course groups (two-way ANOVAS and Schéffe test, n=205)

Course group		1976		1979		1980		1988		ANOVA Source of variance: Sex	Course Group	Inter-action	Schéffe test *) (p<.05)	
		M	W	M	W	M	W	M	W	df=1	df=3	df=3		
		n=26	n=43	n=21	n=32	n=16	n=25	n=21	n=21	F	F	F	M-M	W-W
F1: Teacher's personal congruence/ genuineness (-)	M	-.15	.01	.10	.07	.08	-.13	.01	.01	0.03	.42	.45		
	SD	.61	.85	1.01	.82	.65	.56	.84	.85					
F2:Social form: class-centered (+) - individual – centered (-)	M	-.21	-.23	.25	.01	.28	-.14	.18	.29	1.56	3.66**	1.12	76-88	
	SD	.75	.65	.77	.66	.81	.72	.71	.89					
F3: Teacher involvement (+)	M	-.04	-.22	-.20	-.17	.02	.21	.51	.42	0.04	5.69***	0.76	76-88	79-88
	SD	.69	.77	.75	.74	.73	.68	.95	.63					76-88
F4: Teacher's fact-centeredness (+) – student – centered-ness (-)	M	-.59	.10	-.06	.03	.01	-.18	.51	.24	3.10	7.48***	6.58***	76-88	-
	SD	.68	.60	.70	.55	.46	.47	.95	.48					

Values are means (SD), F, Fisher's F-statistic (variance ratio); *,**,*** p<0.05, 0.01 and 0.001 respectively
*) t-values were calculated after applying Barlett's test for homogeneity of variance (Schéffe 1959), M=men, W=women

(3) Sensitivity of the rating scale

The earlier reported analyses suggest that students' attitudes concerning their "ideal" PE teacher expectations were soft, with the first general dimension associated with contextual factors (course group and gender). To gain more understanding of the measuring instrument and its sensitivity, a discriminant analysis was undertaken using the (contextual) combined two-tailed course group as the dependent criterion variable and rating scale variables as predictors.

Table 53 displays the significant discriminant functions derived from predictor variables: for the male (Wilks' Lambda = .62, df 6, p< .001) for the female (Wilks' Lambda = .57, df 12, p< .001) and for the total sample (Wilks'

Lambda = .68, df 12, p < .001). It might be noted that all of the added items (17, 18, 19 and 20) were selected to the functions in the first steps and their F-ratios were statistically significant at one percent level.

Table 54 highlights the sensitivity, the power of the discriminant functions, to classify contextual combined groups subjects correctly to their own group (1970's/1980's). The discrimination power of the functions was high: overall, 81.1% of the male, 83.5 % of the female and 77.1 % of the total sample were correctly classified. In the female sample the discriminability of the function was more stable in both groups (84 % - 82 %) than that of the male (87 % - 73 %) and of the total sample (77.9 % - 74.7 %). This is also an indication of gender groups' differences concerning their expectations for "ideal" P.E. teacher's characteristics.

TABLE 53 Standardized canonical discriminant function coefficients and univariante F-ratios based on students' attitudes, "ideal" P.E. teacher's characteristics expectation ratings for two-tailed combined gender course groups representing populations before and after the study programme reform (1978)

Variables	Male (n=84)		Female (n=121)			Total (n=205)		
Item no	Function	F-ratio	Item no	Function	F-ratio	Item no	Function	F-ratio
17	.66	15.5***	16	.47	16.33***	17	.33	19.63***
19	-.48	6.48**	18	.66	9.04**	14	.31	14.15***
6	.49	4.14*	14	.51	11.63***	6	.48	9.83**
7	.44	5.47*	9	-.38	5.86*	19	-.34	8.54**
13	.36	5.02*	20	.33	4.33*	18	.38	7.12**
20	.29	2.87	6	.35	5.59*	16	.32	5.30*
			1	-.30	.76	7	.35	7.34**
			19	-.24	2.94	20	.32	7.91**
			10	-.21	4.52*	8	.77	7.91**
			7	.18	1.09	9	-.17	4.24*
			2	.16	.55	1	.17	.87
			17	.15	6.54**	11	.14	3.13
Eigenvalue .62			Eigenvalue .77			Eigenvalue .48		
RC=.62***			RC=.66***			RC=.57***		
Wilks' Lambda .62, df=6			Wilks' Lambda .57, df=12			Wilks' Lambda .68, df=12		
Chi Square 38.16, ***,			Chi Square 64.38***			Chi Square 76.97***		
*=p<0.05, **p<0.01, ***=p<0.001								

TABLE 54 Summary of discriminant function analyses to classify the decade groups (1970s and 1980s), in-group and inter-group by students' "ideal" P.E. teacher's characteristics expectation rating scale variables

Actual group		Number of cases	Predicted group membership 1970	1980	Percent of grouped cases correctly classified
Male	(70)	47	41	6	
			87.2%	12.8%	
	(80)	37	10	27	
			27.0%	73.0%	**81.1%**
Female	(70)	75	63	12	
			84.0%	16.0%	
	(80)	46	8	38	
			17.4%	82.6%	**83.5%**
Total	(70)	122	95	27	
			77.9%	22.1%	
	(80)	83	21	62	
			25.3%	74.7%	**77.1%**

12.5.5 Discussion and conclusions

The aim of this program evaluation study was the validation of the measuring instrument, "ideal" P.E. teacher expectations rating scale in the evaluation of the study unit program. It is an adaptation of the bipolar 16 item, 1-6 scale, instrument developed and validated by Hytönen & Komulainen (1971) based on ideas of Flanders (1965, 1970) and Rogers (1967). The scale was considered applicable in the framework of this study, to be used as an intervention, as a means for enhancing students' goal orientation and learning in the course of didactic observation and microteaching and to be used for diagnosing student's attitudes before the beginning of the training sessions. Based on the framework of study program four - items, connected to the social form, division of the labour and responsibility and teachers own motor engagement - were added (Heinilä 1988, 1992a).

The reliability of the questionnaire, in terms of homogeneity, was estimated by computing Cronbach's alpha, on the basis of item - test correlation for students in four subpopulation groups (n=194). The median value was .50 (.43 - .71) and, on the basis of four factors solution among four subpopulations, it ranged from .56 to .92 and was deemed to be sufficient for further analyses. The reliability of the measuring instrument was also demonstrated in the subsequent validation studies.

The validity, in terms of construct validity, was assessed by using a four factor varimax-factors solution and by determining in-group and out-group variance of subscales in subpopulation groups based on a sample of 30 % the study unit population (1976 – 1988) and with 100% response rate. Firstly, the findings paralleled with the results obtained in an earlier study conducted by Hytönen & Komulainen (1971) concerning the construct and the factor structure and the general bipolar factor dimension (1). It explained one half of the common variance in both studies and it loaded on the same items called the "student-centered–teacher-centered" teaching style in the earlier study; in this study it was named *"teacher's congruence/genuineness"*. It was found to be consistent over decade contextual group and gender variation. Also factor three loaded on the same items as did factor two in the earlier study, whereas the second factor was unique - connected to "Social form" - and loaded on the added items. The structure was clear based on the low inter-correlations between factors (r = .02 – r = .19) and it was found to be stable by comparing the different subpopulation groups of the 1980 sample (n= 83).

The correlation between subjects' attitudes, "ideal" P.E. teacher expectation general factor (I) ("teacher's congruence/genuineness (-)") scores and student's intake test scores measured two years earlier and based on the 1980's population data were r = .48 in the practice test and r = .58 in the total scores and significant at one percent level. They were logical indicators of the stability of students' entry characteristics, attitudes and motivation, also toward the practice study unit programs (Heinilä 1988). The results are congruent with the basic theory of Flanders (1965) and Rogers (1967). The result obviously also support the findings of earlier studies (Cloes, Hilbert & Piéron 1995, Henry 1969, Hytönen 1973, Whitehead 1980). Moreover, the correlations between

subject's attitudes, their expectations in subscale three (F3), "teacher involvement" and their own entry teaching behavior sum scores measured simultaneously (n = 42) was high (r = .40) and statistically significant at one percent level as assumed. This finding is connected to the stability of personality and congruent with results obtained in studies conducted e.g. by Hanke (1980b), Heinilä (1992), Hytönen (1973), Kane (1968) and Siedentop (1981). It also indicates a good reliability and validity of the measurement used in both tests connected to the microteaching course program. Also the inter group and between group variation finding in the data gathered before and after study program reform at the faculty (1978) was an indication of the relationship between the program environmental setting and subjects' presage variables, as assumed. This, in turn, also indicated the good reliability and external validity of the program element in the measuring instrument used. Moreover, the results of the discriminant function analysis indicated that the measurements used for determining sensitivity of the rating scale among two tailed subpopulations appeared to be adequate given the level of discriminant validity in that 81%, 84% and 77% of the male, female and total sample subjects being correctly classified into their own combined decade course groups (i.e. in-group or out-group) on the basis to their "ideal" P.E. teacher characteristics expectations. In addition, the results of these analyses indicated the good validity of the 20 bipolar items questionnaire with four added item, social form and movement issues.

It can be concluded that these findings demonstrated the conceptual integrity and coherence of the instrument with the framework of the study unit program used as a means of intervention and supported the claim of at least a satisfactory level at validity and sensitivity. Nevertheless, given the complexity of attitude formation, it is recognized that other factors beyond those presented in this investigation (e.g. Telama et al 1988, Telama 1990) may also be important contributors to attitudes, and to the study of motivation at the faculty longitudinally. Further research into interventions designed to develop attitudes and intentionality of student teachers, within the preservice setting is warranted (see also Locke 1986, Martin et al. 2001).

Based on the results, it was judged that the validity of the measuring instrument was sufficiently high to be used in further analyses, e.g. for the assessment of programs internal and external predictive validity in contextual variation, and also to be used as a means of intervention in the course of didactic observation and microteaching.

13 PHASE III: PROGRAM PREDICTIVE VALIDATION, A MULTIVARIATE LONG-TERM APPROACH

13.1 Introduction

13.1.1 Background and purpose

This substudy arose out of a conviction, based on observation and research reports –referred the earlier phases of Section 1 and 2 – connected to P.E. teacher training program evaluation, that in general they lack reliability and validity (e.g. Clarce 1971, Melograno 1971, 1979, 1985, Lawson 1988, Locke 1983, Siedentop 1986, Silvennoinen et al. 1991). Evaluation of teacher training programs aimed at the acquisition of teaching skills have shown that subject personal characteristics are related to the effects of experimental treatments and training programs intervention strategy – learning of teaching behavior (e.g. Hanke 1980b, Siedentop 1981); secondly, that there have not been found significant changes in personality e.g. across three years in teacher training (e.g. Hanke 1980b, Hytönen 1973, Kane 1968).

Based on previous studies considering problems connected to program evaluation in colleges, as in the Faculty of the Health and Physical Education at the University of Jyväskylä (Heinilä 1988, Laakso 1984, Rantakari & Tiainen 1983; Silvennoinen et al. 1991, Telama 1967, 1968, Telama et al. 1988) – where the study object course was conducted from the year 1974 onward (-1991) in different kind of social settings, before and after the study degree program reform (1978), and changes made in students selection procedures – *it was assumed that these contextual factors are related to teacher training pre-interactive and interactive process and output – and reflect also to the criteria used in the assessment of the predictive validity of an educational program.*

The inter individual variability indicated in studies as well as contextual variation related to subjects' presage variables, and content, process and outcome variables underlines the necessity of replicated investigations

concerning the effects of contextual variables (Dunkin 1987). Such continuous measurement systems will permit observation of how curriculum packages overlap in procedures, as well as ways and extent in which they really differ. Each study unit is assumed to be unique, and e.g. the effects of the study degree program reform for assessing of the validity criteria for different study units also need to be studied separately.

The purpose of this particular study was (1) to describe the importance of selected student teachers' personality characteristics in determining variance in teaching behavior in the microteaching course and variance of success on the didactic observation and microteaching course and (2) to investigate to what extent variance in student process/behavior and success in didactic observation and microteaching course is accounted for by contextual variables, such as by student selection procedures and changes in the curriculum (the study degree program reform (1978) of the Faculty of Health and Physical Education, University of Jyväskylä) and (3) to describe students' program evaluation in contextual variation.

13.1.2 The study unit: course of didactic observation and microteaching

The course package was based on the theory of Flanders (1965, 1970) with the objective to study the knowledge and mastery as well as cognitive understanding of characteristics of teacher-student interaction as defined by the author's adaptation of Flanders' Interaction Analysis System, PEIAC/LH-75 I and PEIAC/LH-75II, (Heinilä 1977a, 1977b). The program, its intervention strategy model, main components, methodology, as well the assessment of the validity of its basic elements have been reported in earlier phase (II) A-D. The contextual frame of the program, i.e. the study degree program reform (1978) and student intake procedure at the faculty were discussed in the earlier phases in sections 1 and 2, introduction and in connection with earlier studies (Heinilä 1988, Telama 1975, 1979).

13.2 Research task

This study sought answers to following research questions concerning the sample, study unit populations, the course of didactic observation and microteaching:

1. What factors account for predictability of students' study success on the study unit – a course of didactic observation and microteaching?
2. What differences exist in the extent of predictability of students' study success between the male and female students?
3. What differences exist in the extent of predictability of students' study success between the different intake-course groups?
4. What differences exist in predictability between the intake-course populations before and after the study degree program reform at the faculty?

13.3 Methods

13.3.1 Sampling design

Based on previous studies (Heinilä 1988) and on frameworks presented in the related literature (Dunkin 1987, Pedhadzur 1982) a regression model was designed in which the relationships of the predictor and criterion variables were studied. The interest is more in the criterion, thus students' achievements as the predictor variables, and the objectives of this study are concerned with prediction and explanation. The design of this multidimensional ex-post-facto program evaluation inquiry is an application of Dunkin's model (1987). The framework, its components in relation to other components and research strategy are presented in Figure 34 and the research design in Figure 35.

In this design each measurement taken on a subject at a particular point of time is assumed to be influenced by three contextual factors: (1) subjects background (as sex and prior school success and prior experiences); (2) students selection procedure and intake course group; and (3) study unit programs' contextual setting in the faculty before and after the study degree program reform at the faculty. Thus, the problems of these designs are firstly, cross-sectional confounded by subjects' background and course group effects; secondly, longitudinal, confounded by students' back-ground and time effects, and because measurements are made at different times of the study period (intake-course, 2^{nd}-3^{rd} study-year) study unit/effects in the time-lag designs are confounded by course group in contextual setting and time measurement effects. To study these confounding effects separately, a multivariate longitudinal design is used which means that investigations are repeated in more than one intake course group with overlapping in study periods in the faculty and in study object unit and time of measurements.

13.3.2 Procedures

The data of this particular study cover several successive student intake course groups during 1974-1988. The subjects participated in a didactic observation and microteaching course as second and third –years students. As Gage and Berliner (1979, 699) suggest, "to estimate the criterion validity we need to test a group of students and let all of them, regardless of score, come into programme".

The sample used in this study was comprised of intake/course populations 1974/1976, 1976/1979, 1977/1980, 1986/1988 is representing 30.3% of the study unit population. As the criterion for sample were used participation in all intake tests and in all interventions, tests in the course of didactic observation and microteaching and criterion performance on the course. The males made up 41% and the female 59% of the sample, n=205. The dropout in the four-intake course population varied between 9-19% of the intake population (Appendix 9).

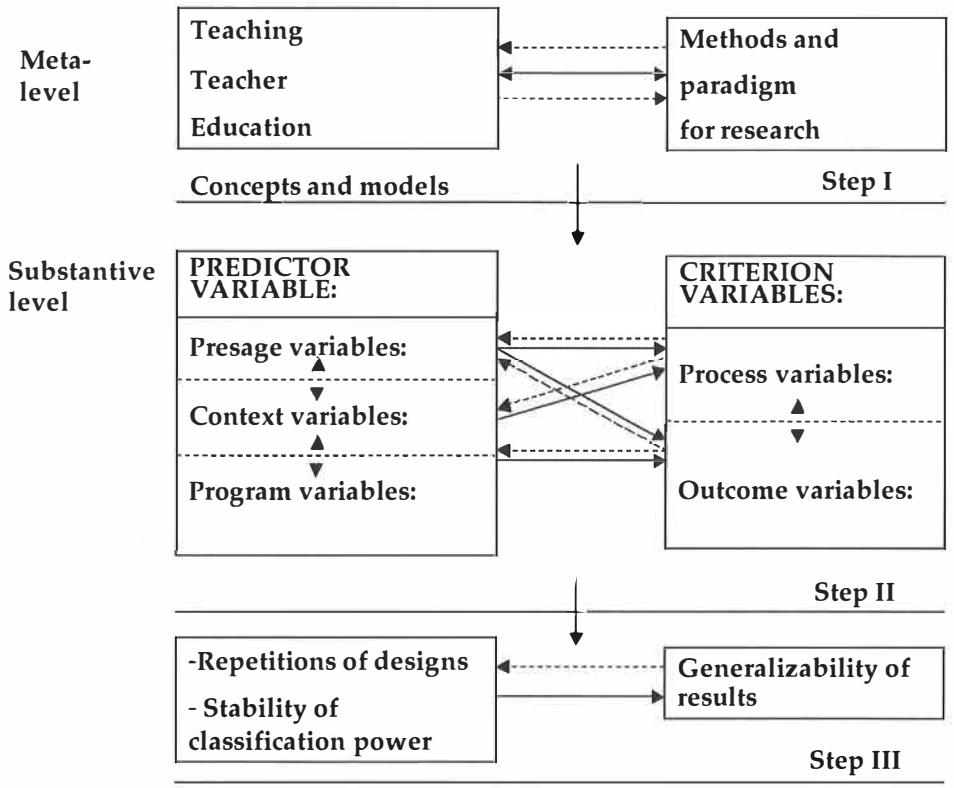


FIGURE 34 Research strategy; schematic representation of sections in relation to other sections and assumptions of the study (Heinilä 1992a)

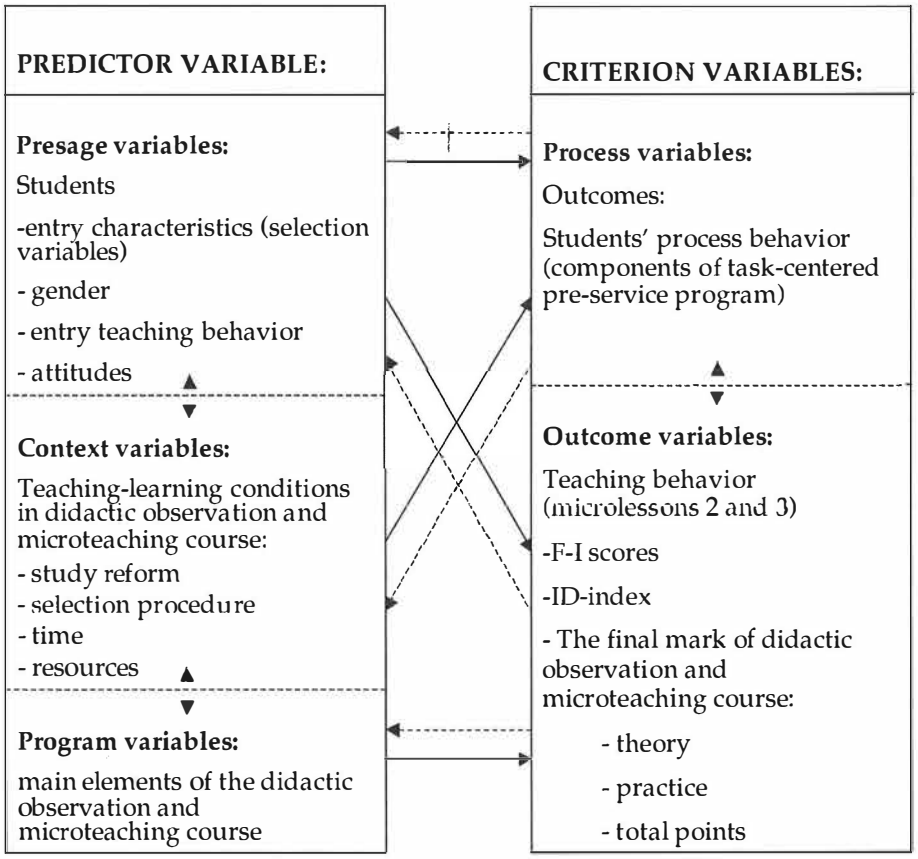


FIGURE 35 Research design

13.3.3 Variables and instrumentation

(1) Predictor variables

The selection variables were: stage I: sum scores (prior school success), stage II: theory test scores (weight 30.7%), practice test scores (weight 55.5%), rated teaching episode (weight 11.5%). They were tested by the teachers of the faculty and coded from the selection protocol documents of the faculty. These weights were used in the entrance examination after the study degree program reform (Appendix 9.1).

The program intervention variables: Students' measured entry teaching behavior (control) was observed by using PEIAC/LH-75, II - a multidimensional observation instrument (cluster I verbal, II movement), 16 and 2 categories, 6 sec time unit. Reports on research methods (e.g. instrument, data gathering, and reliability and validity) have been published separately (Heinilä 1977b, 1980,

1987, 1988, 1990). (See Phase II A and B) In the case study I A, altogether 126 microlessons (5-10 min, 12600 6 sec. time units) were observed and recorded from video taped material by a trained observer. Reliability-objectivity of coding (13 cases, 1300 6 sec. time units) was estimated using Scott's Pi coefficient between the researcher and the observer. Inter-coder agreement: cluster I .76, II .84; within-coder constancy: I .98, II .98; between-coder constancy: I .73, II .98.

Students' rated entry teaching behavior (control): rating scale of 4 items (clarity of presentation, understanding of task content, teacher-pupil interaction, creativity, and 1-6 scale. Reliability between two judges (1988) observing video recorded control lessons (5 min), 13 cases, 126 min., r .36 ($p < .01$); Kendall's W ranged between .56-70 ($p < .05$). Validity: stability of teaching episode rating during the selection procedure and two years later on entry to the course (n=42), r .52 ($p < .01$) (see Phase II C, Appendix 7.1).

Students' attitudes: Personal expectations concerning "ideal" P.E. teacher characteristics; questionnaire, 20 items, scale 1-6; based on ideas of Flanders (1965) and Rogers (1967). (Appendix 8.1). *Validity* of the original instrument was demonstrated in previous studies (Hytönen & Komulainen 1971) and of its application by author (Heinilä 1988, 1992a, 2001) (see Phase II D). Reliability of the questionnaire in terms of the homogeneity, Cronbach's alpha, was based on four factor solution correlations among four populations and it ranged from .56 to .92 (see Phase II D, Appendix 8.2).

(2) Criterion variables

The five selected criteria were (1) students' teaching behavior (2nd and 3rd microlesson (PEIAC/LH-75 II) F-1 scores "teacher initiation" (-) vs. response behavior (+)", (2) ID-index, "non-directive teaching" (2nd and 3rd microlesson) taken from the validation study (Phase II B), (3) the final mark of the theory section, (4) the final mark of practice and (5) the final note of the course (theory and practice) from the combined course of didactic observation and microteaching (1978-); taken from the documents of the Faculty (scale 1-3; <2=low level, > 2=high level) (Appendix 9).

13.3.4 Statistical procedures

For the purpose of this study the use of multiple linear stepwise regression analysis and discriminant analysis was deemed to be applicable (e.g. by Dillon & Goldstein 1984). The assumptions concerning these methods were controlled by using multiple comparison procedures connected with these analyses: S-method, Schéffe (1959, 73-77), after Barlett's test of homogeneity of the variance (1937). Zero-order correlation coefficients among variables were examined for possible instances of multicollinearity. If multicollinearity existed ($r > .70$), then the variables with lowest zero-order correlation with the dependent criterion variable were dropped from the analysis. Using this criterion, the intake sum scores, didactic observation course sum scours and entry teaching behavior sum scores were dropped from the equation because of their high correlation

with the other intake test and other entry teaching behavior scores, and because of their high correlation with the other intervention variables (Appendix 9.4).

A discriminant function was calculated for each cohort to analyze the strength of the association of each measure with success MD-group for the male, female and total population. The form of the discriminant function was the same as in stepwise regression analysis. With the help of this function it was seen how many percent of students were correctly classified with respect to their MD-groups (low-high achievement). The probability level of $p < .05$ was initially set as an indicator of statistical significance in all analyses.

13.4 Case study I A: Prediction success in student teaching from students' selection variables, rated and measured teaching behavior and students' attitudes

13.4.1 Results

(1) Relationships between predictor and criterion variables

Results concerning variable frequencies and correlations are presented in Appendix 9. The results obtained in the first step of this ex post factor inquiry controlling the assessment homogeneity of the variance of the predictor and three criterion variables by t-test, revealed that the male (n = 21) and female (n = 21) population were statistically significantly different in the first stage of student selection (prior school success sum scores, $t = -3.26$, $p < .002$) and in the second stage in teaching episode sum scores $t(42) = 2.53$, $p < .01$) as well as in the entry teaching behavior (sum scores of teaching episode two years later control, $t(42) = 2.72$, $p < .01$). Also total achievement scores on the course were higher for the females ($t(42) = -2.51$, $p < .02$). Thus the female population represented a higher and more stable cognitive level whereas the male students were rated higher in teaching behavior both in the selection in teaching episode ($t(42) = 2.52$, $p < .01$) and control before the course ($t(42) = 2.52$, $p < .01$) as well as in the final measured assessment. However, it was identified that the female population was more consistent e.g. in rated teaching behavior: statistically significant Pearson's correlation coefficients were found between teaching episode sum scores during the selection stage two and the two years later control lesson was .65 and significant at the .01 level of confidence, and between the final mark of the course .34 significant the .05 level whereas with the males these correlations were lower: .33 and -.09. The course sample (n = 42) correlations between the following predictor variables and criterion (the final note) were statistically significant: the intake stage I, school success, $r = .26$, $p < .05$; stage II theory test scores $r = .35$, $p < .01$ and gender $r = .37$, $p < .01$. Based on these findings it was not deemed reasonable to use gender as a predictor, but analyse the gender populations separately and compare the results of these replicated designs. The results of all regression analyses are summarized in

Appendix 9.5.1 - 9.5.9. The correlations between the predictor and criterion (1, 2 – 5) variables in case study IIIA can be found in Table 55.

TABLE 55 Pearson's correlation coefficients between predictors and criterion variables. (1) students teaching behavior (mean of microlesson 2 and 3) (2) F-1 score, (2) ID-index and (5) the final mark of the course for male and female, course 1988, n = 42

	Course 1988					
	Men (n=21)			Women (n=21)		
Variables	(1)	(2)	(5)	(1)	(2)	(5)
STUDENTS' ENTER CHARACTERISTICS SELECTION PROCEDURE:						
(1) stage I: sum score	-0.08	-0.09	-0.23	-0.35	-0.44	0.47*
stage II:						
(2) theory test score	0.23	-0.04	0.38*	0.16	0.29	0.22
(3) practice test sum score	0.38*	0.24	0.14	-0.22	-0.17	0.16
(4) teaching episode sum score	0.42*	0.35	-0.09	0.20	0.34	0.34
(the total score of the selection procedure)	(0.41*)	(0.19)	(0.28)	(0.00)	(0.17)	(0.30)
(7) STUDENTS' ATTITUDES REFLECTATIONS CONCERNING "IDEAL" PE-TEACHER CHARACTERISTICS:						
F1 Teacher's congruence/ genuineness (-)	-0.19	-0.26	-0.54**	-0.12	-0.28	0.03
F2 Social form: class-centered (+) - individual centered (-)	-0.14	-0.12	-0.10	0.20	0.24	-0.13
F3 Teacher involvement (+)	-0.11	0.02	0.01	0.41*	0.43*	-0.19
F4 Teacher fact-centeredness (+) vs. student centeredness (-)	0.19	0.08	0.02	-0.18	-0.22	-0.20
(5) STUDENT TEACHING BEHAVIOR (CONTROL): Rated teaching episode 5 min:						
Item 1: presentation	0.11	0.26	0.10	0.45*	0.54**	-0.41*
Item 2: understanding of task content	0.24	0.40*	0.14	-0.08	0.16	0.07
Item 3: teacher-pupil interaction	-0.13	0.11	0.29	0.08	0.34	0.14
Item 4: creativity	0.14	0.26	0.34	0.09	0.25	0.19
(sum score)	(0.13)	(0.30)	(0.27)	(0.15)	(0.40)	(0.19)
6 STUDENTS' PROCESS-BEHAVIOR (CONTROL) F1: teacher initiation (-) vs. response behavior (+)	0.01	0.14	0.06	0.36	0.27	0.28
(n) ID-index	-0.10	0.12	0.01	0.15	-0.04	0.30

* = significant at the 5% level (2-tailed)
** = significant at the 1% level (2-tailed)
( ) = not used in multiple regression analysis

(2) Results of the regression analyses

When regression analysis was run on the intake-course 1988 *male student* population, the following picture presented in Tables 56, 57 and 59 emerged:

(1) *When the students' measured teaching behavior (of 2nd and 3rd lesson) F-1 scores, "teacher initiation (-) vs. response behavior (+) was the criterion of achieving success,* a three variable model was extracted: two selection variables related to teaching episode scores accounted for 18%, theory test scores for 11% and additionally student attitudes (F-1 scores, "teacher congruence genuineness (-)") for 16% (total of explained variation 45%, F (3,17)=4.63, p< .01). The discriminant power of this variable combination between low and high achievement groups (<2>2) was 67% and statistically significant at the .05 level of confidence.

(2) *When the second measured student teaching behavior variable - ID-index "non-directive teaching" (of 2nd and 3rd lesson) - was used as the criterion of study success,* a five variable model was extracted, in which the students' entry teaching behavior variables (control), measured ID-index accounted for 15% and rated teaching behavior (control)/ item 1, clarity of presentation for 8%/ item 2, understanding of task content, for 16% and item 3, teacher-pupil interaction for 15%. Also students' attitude F-1 scores ("teachers' congruence/genuineness (-)") accounted for 12% (total of explained variation 66%, F(5,15)=5.83, p< .001). The classification power of these variables was 86% (p< .01).

(3) *When the final mark of theory* was used as the criterion, a one - variable model was extracted for the male students: intake theory test accounted for 18% of the variation (F (1,19=4.13, p< .05) and its classification power was 67%.

(4) *When the final mark of practice* was used as the criterion, also a one-variable model was extracted, where students' attitudes, F1 scores "teacher congruence/genuineness" accounted for 35% (F (1,19)–9.98, p< .001) of the variation and its classification power was 67% (p< .01)

(5) *When the final mark (theory and practice) of the course was used as the criterion variable,* a two-variable model was extracted: students' attitude, F1 scores "teachers' congruence/genuineness (-) accounted for 29%, students' rated entrY teaching behavior (control)/ item 4 (creativity) accounted for 16% (total of 45%, F (2,18)=7.35, p< .01). The classification power of this variable combination was 71% (p< .08) (criterion 3 – 5, see Table 56).

TABLE 56 Results of the regression analyses for the male student intake course 86/88 (n = 21). Regression coefficient b, standard errors in brackets and standardized regression coefficients ß

Predictor variables	**Criterion variables**					
	Theory scores		Practice scores		Final mark	
	b	β	b	β	b	β
Selection variable						
Stage II - theory test scores	.28(.14)	0.42*				
Students' rated entry teaching behavior (control) - item 4, creativity					1.40(.62)	0.40*
Students' attitudes expectations concerning characteristics of "Ideal" P.E. teacher: Factor I scores: "Teachers' congruence/ genuineness (-)"			-3.21(1.02)	-0.59**	3.46(1.10)	-0.58**
Constant	9.15(5.25)		21.24(.83)		15.40(2.39)	
R	0.42		0.59		0.67	
R^2	0.18		0.34		0.45	
F	4.13*		9.98**		7.35**	
Classification power	67%*		62%		71%	

* = p<.05
** = p<.01

Thus, the selection variables, stage II (theory test and teaching episode scores) as well as students' attitudes and rated entry teaching behavior (creativity) were important stable predictors for the males. The predictability of students' achievements in practice and especially in measured teaching behavior, ID-index (the non-directive teaching) was higher than of the final theory mark and final course mark (see Tables 58 and 59).

For the female students, using the same predictors and criterion, the following models were extracted:

1. *Teaching behavior, F-1 scores:* The only variable selected was students' rated entry teaching behavior (control), item 1 (clarity of presentation), accounted for 21% of the variation: (F (1,19)=9.49, p< .05) and its classification power was 62%.
2. *Teaching behavior, ID-index*, the same variable students' entry teaching behavior, item 1 as above, accounted for 30% of the variation: (F (1,19) =8.94, p< .01) and its classification power was 81%, p< .01 (see Table 59).
3. *The final mark of theory*: intake prior school success accounted for 33% of the variation (F (1,19)=9.46, p< .01) and its classification power was 81% (p< .01).

4. *The final mark of practice*: A three-variable model was extracted: prior school success accounted for 28%, entry teaching behavior (control) item 1 clarity of presentation for 14%, and item 4 (creativity) for 10%; total of explained variation was 63% (F (4,16)=6.88, p< .01), and the classification power was 81% (p< .01).
5. *The final mark of the course*: the selection variables, prior school success accounted for 22%, and rated entry teaching behavior (control), item 4, and creativity for 12%, adding up to a total of 34% of explained variation (F (2,18)=4.56, p< .01). Its classification power was 76% (p< .05). Table 57 contains the results of these analyses, 3 – 5.

TABLE 57 Results of the regression analyses for the female student intake course 86/88 (n = 21). Regression coefficient b, standard errors in brackets and standardized regression coefficients ß

Predictor variables	**Criterion variables**					
	Theory scores		Practice scores		Final mark	
	b	β	b	β	b	β
Selection variable						
Stage I						
- school success	1.06(0.34)	0.58**	1.58(0.33)	0.90***	1.07(.37)	.57**
Stage II						
- theory test scores			.25(0.13)	0.30		
Students' rated entry teaching behavior (control)						
- item 1, clarity of presentation			2.92(1.20)	0.46*		
- item 4, creativity			1.66(0.81)	0.33*	1.90(1.09)	.35
Constant	-11.51(11.17)		-54.29(15.21)		-17.58(13.51)	
R	.58		.80		.58	
R^2	.33		.63		.34	
F	9.46**		6.88**		4.56*	
Classification power	71%*		81%**		76%*	

* = p<.05
** = p<.01
*** = p<.001

These results indicated that the cognitive components such as the prior school achievements and students' entry teaching behavior, clarity of presentation and creativity were the important stable predictors for the female students' study success on the course of didactic observation and microteaching.

(3) Comparison of predictability by gender

To answer to the second research question: "What differences exist in the extent of predictability of study success between the male and female students?" – a comparison, based on results reported earlier above in tables 56 and 57 and in tables 58 and 59 was conducted, concerning mainly the differences identified in the power of predictability in the two main criterion variables (1) *in the final mark (theory and practice) of the course* and, (2) *in teaching behavior, (ID-index 2nd and 3rd microlesson) (PEIAC/LH – 75 II).*

The predicatability differences in the final mark were visible both in content and in power of prediction models (R^2). For the *male* affective component and study motivation accounted for two thirds (29%) of the total explained variance (45%) combined with mastery of teaching behavior; whereas for the female the cognitive component (prior school/success) explained two third (22%) of the total explained variance (34%) in combination with the mastery of teaching behavior, creativity (as with the males). *For the total sample, by using the final mark (theory and practise) as criterion,* the following picture emerged: a four variable model was extracted: intake stage 1.sum scores accounted for 8 %, intake stage 2. theory test scores for 12 %, attitudes, students "ideal" P.E. teacher characteristics expectations (F1) for 7% and student's entry teaching behavior (control), item 4/ creativity for 7 %, adding up to a total of 35 % (F (4, 37) = 4.86, p=.003).

TABLE 58 Summary of discriminant function analyses and classification power: percent of grouped cases (low-high achievement) *in final mark* correctly classified by selected regression model (R^2) variables for the male, female and total sample of intake course 1986/1988

Variables	Correlation with function Male r	Female r	Total r		Correctly classified Male %	Female %	Total %
Students' entry characteristics: **Selection procedure** Stage 1. sum scores (prior school success)	-	.33	.33				
Selection procedure Stage 2. theory test scores	-	-	.69	**low**	**79**	**80**	**68**
Students' attitudes Reflections concerning "ideal" P.E. teachers' characteristic F1 "Teacher's congruence/ genuiness (-)"	-.84	-	-.45	**high**	**57**	**73**	**75**
				overall	**71**	**76**	**71**
Students' teaching behavior (control) item 4, creativity	-.36	.73	-.12				

Male R^2 = .45, F, p < .01; eigen value = .32; RC= .49; Wilks' Lambda = .76; Sig. (df2) = .08 (n = 21)
Female R^2 = .34, F, p < .05; eigen value = .42; RC = .54; Wilks' Lambda = .71; Sig. (df2) = .04 (n = 21)
Total R^2 = .35, F, p = .003; eigen value = .42; RC = .54; Wilks' Lambda = .71; Sig. (df3) = .004 (n = 42)

Table 58 reports significant discriminant functions derived from selected regression model (R^2) variables for the intake course 1986/1988 male (Wilks' Lambda = .76, p = .08), female (Wilks' Lambda = .71, p = .04) and for the total sample (Wilks' Lambda .71, p = .004). Analysis of the correlations between the discriminating variables and the functions revealed also, which variables contributed to the prediction of study success in the teacher-training course for the male, female and total population. The classification table included in Table 58 highlights also the sensitivity of the discriminant function to predict study success (in the final mark) for the subpopulations: overall, 71% of the total sample were correctly classified including 68 % and 75 % of subjects with low and high achievement scores, respectively. For the male sample, the classification power of determinant function was at the same level (71%), but its variation between the low and high achievement MD-group was great: 79 % and 57%. For the females, the classification power was identified to be stronger and more stable than for the males: overall, 76%, of the sample including 80% and 73% of subjects with low and high achievement scores was correctly classified. Thus, the study success was among both gender more predictable for the low than for the high achievement MD-group subjects and especially among the male sample. For the total sample, however, the situation was the opposite: the functions' discriminability was identified to be stronger in the high than in the low achievement MD-group subject classification (75% and 68%).

TABLE 59 Summary of discriminant function analyses and classification power: percent of grouped cases (low-high achievement) *in teaching behavior*, ID-index sum of 2nd and 3rd microlesson correctly classified by selected regression model (R^2) variables for the male, female and total intake course 1986/1988

Variables	Correlation with function Male r	Female r	Total r		Correctly classified Male %	Female %	Total %
Students' entry characteristics: **Selection procedure** Stage 2. teaching behavior sum scores total	-	-	.54				
Students' entry teaching behavior (control) item 1 "clarity of presentation"	.18	1.00	.79	**low**	**100**	**82**	**57**
item 2, understanding of task content	.37	-	-				
item 3, teacher pupil interaction	.11	-	-	**high**	**70**	**80**	**76**
Students' entry teaching behavior (control) ID-index	.15	-	-				
Students' attitudes Reflections concerning "ideal" P.E. teachers characteristic F1 "Teacher's congruence/ genuiness (-)"	-.14	-	-.35	**overall**	**86**	**81**	**67**

Male R^2 = .66, F, p < .01; eigen value = 1.48; RC= .77; Wilks' Lambda = .40; Sig. (df5) = .01 (n = 21)
Female R^2 = .30, F, p < .01; eigen value = .67; RC = .63; Wilks' Lambda = .60; Sig. (df1) = .002 (n = 21)
Total R^2 = .30, F, p = .003; eigen value = .26; RC = .46; Wilks' Lambda = .79; Sig. (df3) = .03 (n = 42)

(2) The second important criterion in the assessment of the predictive validity of the program *was ID-index, indicator of students' non-directive teaching behavior learning gain.* Table 59 reports significant discriminant functions derived from the selected regression model (R^2) variables for the intake course 1986/1988 male (Wilks' Lambda .40, p< .01), female (Wilks' Lambda .60, p = .002) and total sample (Wilks' Lambda .79, p = .003). Analyses of correlations between the discriminating variables and functions revealed also which variables contributed to prediction. *For the total sample (n = 42), a three-variable model was extracted: students' entry teaching behavior, (rated in the student selection procedure stage 2. teaching episode sum scores) and – rated and measureD two years later - in microteaching episode (5 min. control) item 1., clarity of presentation accounted for a major part 8 % and 15% of the total of 30 % explained variance in study success combined with students attitudes, their "ideal" P.E. teacher characteristics expectations (F1, "teachers' congruence/genuineness (-)") (7 %) (R^2 = .30, F = 5.42; df 3.38, p = .003).* The total sample analyses produced highly valid results – significant at .01 percent level of confidence. However, the results shown in Table 59 also revealed great variation in predictability by gender. The classification table included in Table 59 highlights also the sensitivity of discriminant function to predict study success in measured teaching behavior (PEIAC/LH-75 II) – ID-index using regression model (R^2) variables: overall 67 % of the total sample were correctly classified, including 57 % and 76 % of subjects with low and high achievement scores, respectively. For the male sample, the classification power of the determinant function was higher: overall 86 % of the sample were correctly classified including 100 % and 70 % of subjects with low and high achievement scores respectively, whereas for the females the determinant functions' classification power was indentified to be nearly at the same level as for the males (81 %), but it was more stable and sensitive in classification both of the low and high achievement scores subjects: 82 % and 80 %.

Thus, students' entry teaching behavior was found to be a stable, consistent predictor of study success – "art of teaching", learning gain of non directive teaching behavior – combined with the cognitive and affective component. Predictability differed by gender. In this contextual setting, the low achievement (ID-index level) male subjects were highly discriminible (100 %), whereas only 70 % of the total sample, subjects with high achievement could be correctly classified.

(4) Summary, discussion and conclusions, case study IA

This sub-study investigated the relationship between selected student teachers' entry characteristics and their possible mediation of study success in a preservice study unit program, using variables representing teaching behavior achievement and the final marks of the course. It has been pointed out that because of the limitations of ex-post facto inquiry, the conclusions are probabilistic, reducing uncertainty, but not totally eliminating it. However, this must be kept in mind e.g. in comparing predictability by gender, the contextual confounding effects, such as student selection procedure and study degree program reform (1978). In the frame of this case study intake course

1986/1988,the student selection (1986) produced two significantly different gender populations concerning students' cognitive capacity and level of entry teaching behavior when relationships between predictor and criterion variables were analyzed. Also the main criterion, the final mark of the combined course of didactic observation and microteaching was after curriculum degree program reform at the faculty more theory weighted (60,6%) than before. Therefore, the interpretation and explanation of these findings is problematic – context– related (see e.g. Silvennoinen et al. 1991, Whitehead 1980).

The regression analyses for the total sample (n = 42) provited further insight regarding the relative contributions of predictor variables in the final mark. For instance, student selection procedure variable - stage 1.sum scores (prior school success) and stage 2. theory test scores - accounted a major part,(57 %) of the total explained variances (35 %) in study success (final mark, theory and practice). Further variance accounted for by students' study success via attitudes, their "ideal" P.E. teachers' characteristics expectations (F1 scores "teachers' congruence/genuineness (-)"), was equal (7 %) to that attributable to all the four combined (7 %) students' entry teaching behavior scores item 4. creativity. *Predictability of measured teaching behavior, ID-index* was found to be at a higher level than of the other criterion. The most important predictors accounting for 76 % of the total explained variance (R^2 = .35) of the success was students' entry teaching behavior, measured between two years interval, combined with students' attitudes. The classification power of these predictors was for the males 86 %, for the females 81 %, and for the total sample only 67 %. Predictability was, however, more stable for the females for whom "entry teaching behavior, clarity of presentation" was the most important predictor of study success.

Results of regression analyses also showed that student intake practice test sum scores and standardized intake sum scores failed to contribute to the prediction of study success. At all events, the results of this study support the basic assumption concerning effects of the contextual factors on the variation of predictability of students' study success. Data analyses also revealed that the assumption concerning the relationship between student personality, characteristics and teaching behavior was supported, and results also paralleled with the results obtained by Hytönen (1973), Hanke (1980b) and Siedentop (1981) The results also supported the claims concerning the program's internal and external predictive validity. However, for the assessment of program predictive validity a multivariate long-term approach with replicated designs was needed. The summarized results of the other case studies will be presented and analyzed in the next stage.

13.5 Summary and results of replicated case studies: variation of the extent of predictability in study success by criterion and sex among course groups' population

13.5.1 Descriptive information on student background variables by sex and course group

The results obtained in the first step of this study, by controlling homogeneity of the variance of predictor and criterion variables among the four intake course and combined course groups 1. (1970') and 2. (1980') male and female samples by one way ANOVAs, and multiple range test (Schéffe, $p < .05$), are presented in Appendix 9.

Following significant differences between subpopulation groups in student intake and intervention (control) variables and final marks were identified:

1. *The intake stage 1 sum scores* (prior school success) between the males and females differed significantly in two intake courses: 1976/1979, ($t(53)=2.70$, $p < .01$) and 1986/1988 $t(42)=3.26$, $p < .001$. Also in the 1986/1988 course, male teaching episode sum scores were higher than of the female ($t(42) = 2.53$, $p < 01$). Thus, the case study IIIA population was exceptional (control) concerning gender and/or presage variables differences.
2. *Students' rated entry teaching behavior* scores (control) differed significantly ($p < .05$) between the following course groups: 1976-1979, (item 1 and 3) and 1979-1988 sum scores. The scores were higher in the intake course sample, (1976/1977 and 1977/1979) when e.g. teaching episode was weighted in student selection total scores by 25%, whereas after the study degree reform only by 11.5%. However, the male students' rated teaching behavior sum scores were found to be higher after change made when comparing results between combined course groups 1 (1970's) and 2 (1980's): ($t(84)=-3.1$, $p < .01$). (See Appendix 9.)
3. *By comparing the final marks of the combined curriculum course* sample 1 (1970's) and 2 (1980's), significant difference was found ($t(122,83) = 2,83$, $p < .01$), and also between the final practice scores ($t(122,83) = 2.37$, $p < .05$). The female students' achievements level in practice was lower, in the later (2^{nd}) sample ($t(121)=2.0$, $p < .05$) and there were not found any significant differences between the male and female achievement level as before.
4. *In comparing the final marks (theory and practice) between the four intake course sample gender groups,* only one significant ($p < .05$) difference was identified. The achievement level of the intake course 1988 males was lower than that of the 1976. Obviously contextual effects such as the curriculum reform of the faculty were involved in these findings.

13.5.2 Relationships between the predictor and criterion variables by course group and sex

The analysis of relationships between the predictor and the criterion among gender and course groups was based on both zero-order correlations and on stepwise regression analysis and the results are presented in Appendix 9.3.1 – 9.3.3. The data revealed e.g. following important facts concerning contextual variables: firstly, *the total standardized intake scores correlated positively with the final mark of the course of didactic observation and microteaching only in two male groups (1976/1979, r = .24 and 1986/1988, r =.28) and two female population (1976/1979 r = .10 and 1986/1988, r = .30), but these correlations were, however, not statistically significant* (Appendix 9.3.1).

Secondly, the analyses revealed that the student intake, stage 2. *teaching episode sum scores* correlated positively with the total standardized intake scores in four male groups (1974 r = .22, 1976 r = .46, p<.05, 1977 r = .44 and 1986 r = .19), and in three female course populations (1976 r = .28, 1977 r = .22 and 1986 r = .38), but the correlations *were statistically significant only in one male population (1976/1978).* Thus, it can be noted that it statistically speaking depended on chance and on other correlating test scores if students on the basis of the teaching episode test were selected to the Department of Physical Education of the University of Jyväskylä (Heinilä 1988) (see Appendix 9.3.2).

The data indicated, as expected, that the final marks of students´ performances in the combined (1978 -) study unit course of didactic observation and microteaching correlated positively. However, it might noted that the aim and content of course of didactic observation were more theory and science oriented, whereas the course of microteaching was more oriented towards, integration of theory and teaching practice. Also correlations between these two course scores, presented in Appendix 9.3.3, revealed differences between the gender and course populations. *The correlation varied as follows between males (m.) and females (f.) among the four course groups*: 1976, m. r = .69 (p< .01) – f. r = .33 (p< .05); 1979, m. r = .35 – f. r = .09; 1980, m. r = .18 – f. r = .94 (p< .01); 1988 m. r = .58 (p< .01) – f. r = .64 (p< .01). A comparison of correlations between the two combined course group (1. 1970′ and 2. 1980′) populations revealed that the level of correlation was higher in the second groups and any higher among the female populations: m. 1-2 r = .29 – r = .64 (p< .001), f. 1-2 r = .12 – r = .90 (p< .001). This is a clear indication of the contextual effects such as of the changes made to student selection procedure and the degree program reform at the faculty (1978), when e.g. the final mark of the course of didactic observation and microteaching were combined. These contextual facts may have affected the results of this study, as assumed (Appendix 9.4).

13.5.3 Results of regression analyses: comparison of regression coefficients (R^2) and classification power by criterion and sex

The main results summarized in table 60 and in Appendix 9, tables 1-8, indicates how the predictability of students' study success in the course of didactic observation and microteaching has varied longitudinally 1976-1988

when using the final mark of theory, the final mark of practice and the final note of the course (theory and practice) as the criterion and when analysing the results by intake-course group and by gender. The criterion used for the assessment of the extent and level of predictability of success has been the higher value of squared correlation (R^2) in multiple stepwise regression analyses and/or median value (MD) of case-studies, the better predictability of the criterion variance; and the better classification power (percent of cases correctly classified to succeed or not in the course) the stronger the predictability. Moreover, the extent of predictability was also assessed on the basis the number of models extracted in this case studies. The determination of the relative importance of predictors was based also on how frequently predictor variables were selected to the regression (R^2) models.

TABLE 60 Summary of results: squared multiple correlations (R^2), percentages of explanation of regression models of study success and classification power: percent of grouped cases (low-high achievement) correctly classified by variables selected by means regression analyses

	CRITERION VARIABLES					
	Theory score		Practice score		Final mark	
	Men	Women	Men	Women	Men	Women
Course group	%	%	%	%	%	%
1976						
n= 26, 43						
R^2	-	-	42	37	30	-
	-	-	(69**)	(65*)	(77**)	-
1979						
n= 21, 32						
R^2	-	50	30	15	36	40
	-	(84***)	(67)	(63)	(71)	(78**)
1980						
n= 16, 25						
R^2	-	38	43	-	26	12
	-	(72)	(81**)	-	(63*)	(68)
1988						
n= 21, 21						
R^2	18	33	34	63	45	34
	(67*)	(71*)	(62)	(81**)	(71)	(76*)

- = regression model not selected
*, **, *** = p<.05, .01, .001 respectively

The overview the results revealed first the same general facts as identified in results obtained in the case study IA, presented in earlier in this dissertation and in earlier publications (Heinilä 1988, 1992a) by comparing the extent of predictability by criterion and by gender. Firstly, the results indicated that variation of regression coefficients in the criterion variables was related to student's background, gender and also to contextual factors: For *the final mark of theory*, the number of prediction models extracted was lower than in the other criterion. For the male students, only one significant prediction model was extracted and it accounted for variance 18%, and its only selected predictor was intake theory test scores, and its classification power was 67% ($p< .05$). Whereas, for the females: three extracted models accounted for (MD) 38% of the total variance and the classification power of these models was 71%, 72% and

81% ($p < .01$). For the female students, the important predictors for theory achievements were intake stage 1. sum scores (prior school success) and students' entry teaching behavior (item 1 clarity of presentation) and in course group models, after study degree program reform, also students' attitudes (F scores "teacher involvement"). This is, however, an indication of criterion validity: theory practice integration.

By using the final mark of practice as criterion, the results indicated that the extent and level of predictability was higher and also more stable in contextual time variation than for the final mark of theory and that there was not a great difference between genders. *The median value of explained variance of study success in seven sub-population groups was 37 % and the classification power of selected variables varied between 62% and 81% and it was higher in later subpopulation groups.* However, the predictor variables selected for the males and females were different: the important predictors *for the males* were students' selection variables intake stage 2. theory test scores and, after the study degree program reform (1978'-) and before it, stage 1. sum scores (prior school success); and intake entry teaching behavior (microteaching episode sum scores) contributed to prediction. Also - measured two years later - entry teaching behavior (microteaching episode control) items 3. teacher-pupil interaction and 4. creativity scores combined with students' attitudes, F-1 scores "teacher congruence/genuineness (-)" were selected to many male prediction models. *For the females,* also intake variables, prior school success and theory test scores combined with entry teaching behavior (control) item 1. "clarity of presentation" and item 2. "understanding of task content" combined with attitudes, F-3 scores "teacher involvement (+)"- after study reform and before F-2 scores, "class-centeredness (-) vs. individual-centeredness (-)" contributed to the prediction of study success in practice final mark. *Thus, the students' personal characteristics, cognitive capacity, attitudes and mastery in teaching behavior were represented in these significant ($p < .05$) extracted prediction models. The variables selected to prediction models represented both student intake and program intervention strategy – variables measured between two years' interval.* It was an indication of the stability of students' characteristics related to study success in practice achievements. Moreover, as stated in connection with case study IIIA presented above, the correlation between criterion variables the practice final mark and measured PEIAC/LH-75 II ID-index ("non-directive teaching behavior") was $r = .65$, statistically significant at one percent level of confidence (Heinilä 1992a), which also supports the claim of the of program's internal predictive validity.

When the final mark of the course (theory and practice) of didactic observation and microteaching was used as criterion, the median value of explained variance (R^2) in the seven sub-population groups was 34% and varied for the males between 26% and 45% and, for the females between 12% and 40%. Thus, based on the comparison of the amount of regression coefficients (R^2) for determination of the study success, it was identified to be of the same level both for the male and the female students (MD 33 % and 34 %). However, the classification power of the discriminating function of selected (R^2) variable combinations was found to be a little higher and more stable (in in-group – out-group, classification) for the females (MD 76 %) than for the males (MD 71 %).

13.6 Summary and main results of replicated case studies, IIIB

This sub-study was designed for program predictive validation. The results of multiple stepwise regression analyses conducted for four-intake course male and female subpopulations (n = 205) supported the basic assumption of the relationship between selected students personal characteristics and study success in that theory based (Flanders 1965, 1970, Heinilä 1977a) preservice teacher training course, with the aim of enhancing of students' non-directive teaching behavior. The results obtained in replicated case studies revealed also that contextual factors such as students' background, gender and course group were related to the predictability of study success in study unit program as assumed. Its level and extent of variation by criterion and gender was assessed: the greatest difference in predictability between male and female students' study success was found by using the *final mark of theory* as the criterion variable. Predictability was higher and more stable for the female subpopulations (MD 38%) than for the male, for whom only one prediction model was extracted. And by using a combined criterion, *theory and practice as the final mark* of the course, the gender difference was visible but not in the amount of explained variance (R^2), which was identified to be for both genders at the same level (MD 33 % and 34 %), whereas the sensitivity or power of the determinant function to classify correctly subjects with low and high achievement scores was stronger and more stable in the female sample (MD 76 %) than for the male (MD 71 %).

The differences of predictability between genders were visible in the form and content of regression models (R^2). The affective component combined with the mastery of teaching behavior was the best predictor of study success for the male students, whereas the cognitive component combined with mastery of teaching behavior (creativity and clarity of presentation) was the most important predictor for the female students. It was also recognized that predictability of study success *concerning the final mark of practice was* at a higher level than of the other criterion (MD for the males 38 %, for the females 37 %) and after study degree program reform (1978) at higher level especially among the female sample (1988, 63 %). Students' selection procedure variables, intake stage 1. sum scores and stage 2. theory test scores as well as the teaching episode sum scores contributed to the prediction of study success on the course prediction model, whereas the intake practice skills tests sum scores and intake total standardized scores did not contribute to prediction, because their entry was not statistically significant.

Moreover, for the main theory based measure (PEIAC/LH-75 II), ID-index (sum scores of 2nd and 3rd microlesson) the predictability level was higher e.g. among for the male sample (66%) than for the other criteria used in this investigation, as observed in results of case study IA (n = 42). The classification power of the discriminant function of the selected regression model variables was strong: overall 86 % of the male and 81 % of the female subjects with low and high achievement scores were correctly classified. The low achievement level subjects were well discernible among the male (100 %) whereas among the

female the sensitivity of the discriminator was more stable (81 %). Additionally, it can be noted that the correlation of the ID-index with practice final note was r = .65 (Heinilä 1988, 1992a). It was assessed in all case studies, but not used as the criterion in this summarizing comparison (because of technical problems)(Heinilä 1988). This finding was, however, an important fact, an indicator of the program's conceptual integrity and coherence and of its congruence with what was being done in the program intervention strategy realization. Teaching behavior is a "unique" variable, "Art of teaching" (see e.g. Flanders 1987). These results are an indication of a good level of the internal predictive criteria validity of the Flanders-based study unit program: in theory-practice integration combined with the affective component, students' attitudes and motivation.

Although these findings appeared to support the program's external and internal predictive validity, the results of longitudinal investigation connected to students' program evaluation will be presented in the next step.

14 STUDENT PROGRAM EVALUATION AND CONTEXTUAL VARIATION, III B

14.1 Introduction

14.1.1 Background and purpose

This sub-study is linked with earlier presented investigation with the main purpose determining the internal and external validity for the study object course program by coordinating results obtained in replicated case studies.

Also this particular study arose out of an assumption, based on the results of relevant curriculum evaluation studies and observations, that the contextual variables exert an influence on the quality of teaching and on program effectiveness and reflect also to students' perceptions of teaching. In accordance with the idea of Dunkin and Biddle, it is assumed that "recognition of context effects would both clarify and bring greater power to results on effective teaching (1974, 41)".

Most evaluation of teaching at the university level has been based on the feedback of students. Students' ratings based on their own experiences have proven also to be the most valid method (Greenwood & Ramagli 1980). This method was used also in the curriculum evaluation of the Faculty (Rantakari & Tiainen 1983, Telama et al. 1988) and in this study, in Phase I (cf. also Heinilä 1977b). As stated in section I, the research problems of program evaluation (design the work to be done and problems to be solved by teachers) are more or less defined by the frame factors – thus by the institutional setting in which teaching takes places. Empirical research on relationships between the institutional framework factors and the teaching practice itself supports these general sociological points (Heinilä 1971, 1974, Lundgren 1972, Parsons 1968); see Section I pp.75, 77-79 and Larsen (1995, 278).

One objective of the course was to train students to observe, analyse and criticize teaching in its pre-interactive, interactive and post-interactive phases as shown in the program package intervention strategy model, and additionally, in the final stage of the course, also to evaluate the study unit program in the

light of the instructional setting at the faculty, on the basis of their own experiences – (see Phase II).

In earlier stage of the study (Phase IIIA), the objective was to find out to what extent selected students' entry characteristics, as measured in students' intake test and program intervention strategy variables, are predictive of subsequent study success in a study unit, teacher preservice/training course of didactic observation and microteaching conducted at the Faculty of Health and Physical Education of the University of Jyväskylä from the year 1974 onward (-1991) and in two different kind of contextual setting: before and after the study degree program reform at the faculty (1978); and the second objective of the study was to investigate to what extent variance in student teaching behavior and success in the course is accounted for by context variables, by changes in curriculum of the faculty and changes in student selection procedures.

The purpose of this sub-study was (1) to describe teaching given in a study unit program course of didactic observation and microteaching on the basis of evaluation by students, and (2) to describe and compare students experiences of problems encountered during studies, in the frame of contextual variation; and (3) to find out to what extent the variance in students' ratings of program implementation dimensions is accounted for by context variables, such as sex and course group in the frame of the curriculum study degree program reform at the faculty and changes made in students selection procedures (1974-1988).

14.2 Research task

The research tasks in this particular sub-study were:

(1) to describe students' ratings in contextual variation
(2) to find out those discriminant functions that best separate the combined criterion groups 1. and 2. from each other among male, female and total population, in other words maximize the between-group variance, in relation to the within-group variance;
(3) determine the classification power of the evaluation functions extracted (percent of grouped cases correctly classified to their MD criterion group) among male, female and total combined population groups
(4) to find out and validate program implementation factors on the basis of the evaluation of students, in replicated case studies
(5) to describe variance factor scores in the contextual frame by sex and course group;
(6) to find out the effects of contextual factors, sex and course groups and also sex and combined course group on variance of students' ratings on the program "implementation" dimensions.

Thus, the study was connected with prediction and explanation.

14.3 Method

14.3.1 Data

The subjects were male (n=113) and female (n=170) P.E. students, sample 46% of the total population of the study unit in the period of 1974-1988 and enrolled at the Faculty in 1974, 1976, 1977, 1979, 1980 and 1986 (n=286), who were participants of the course of didactic observation and microteaching during the first and second term of the third study year, or during the least term of the second and first term of the third study year. They represented study unit populations before (1) and after (2) the study degree program reform (1978). The Faculty students' evaluations of the program were carried out for every course group in a controlled identical situation at the end of the last training sessions in the microteaching course, administrated by course group leaders. The answers were given anonymously.

14.3.2 Procedures and instrumentation

The program was evaluated by students on a questionnaire, presented in Appendix 5.1. It was used already in the first explorative pilot-study (Heinilä 1977b; Phase I, see Appendix 5). The preliminary form consisted of fifty-eight and, after revision, of sixty-five questions or statements related to the framework of the program: (1) to the goals, contents, forms, relevance of study unit in the curriculum program, (2) timing, (3) time reserved for different units, (4) way of carrying it out (5) use of hand-outs and AV-material, (6) students perceptions of they own learning gain and (7) their future plans to use the method. The five-point rating scale was constructed according to a Likert-type method: the evaluative criterion varying from item to item presented as negative and positive statements on the object ranking from "completely disagreement" (1-) to "completely agreement" (-5).

The reliability of the questionnaire, in terms of homogeneity of the variance, Cronbach's alpha, was computed on item test correlations for students in four-population groups (n=197), and also on seven-factor inter-correlations. It ranged from .56 to .92 (Heinilä 1988). The rating variables were examined using t-tests, factor analyses, analyses of variance, and Schéffe multiple-range test in order to determine the differences between the four course groups. In order to investigate whether the students could be correctly classified into their contextual combined criterion groups on the basis of their ratings, a discriminant analysis was conducted. The comparison of students' ratings among the four course groups was done by using a one-way analysis of variance. The relations and differences between program implementation factors and background variables (as sex and course group and combined course groups) were analysed by using two-way ANOVAs.

14.4 Results

14.4.1 Comparison of students´ ratings between the two combined groups before and after the study degree program reform by sex

The analysis was conducted in three stages and the main results are presented separately. In the first stage the two combined curriculum groups (1) and (2) student ratings were compared by using item frequencies and t-tests among the male, female and total (n=283) samples (Appendix 10.5).

An overview of these analyses revealed a great overall variability in significant differences in comparison to results within and between sub-population groups. The quantity of significant item differences varied between gender groups as follows: 13% of the male, 30% of the female and 31% of the total sample ratings of the total 58 items changed. The trend of the changes based on statement frequencies was identified to be from the earlier extreme point of the scale, of complete disagreement (1) and complete agreement (6), toward the center of the scale. One example of this trend was item no. 11, which reads: "The course has overlapped unnecessarily with my earlier studies": and the mean rating differed as follows for the male 2.0 - 2.5 ($df = 111$, $p < .01$) for the female 1.7 - 2.2 ($df = 168$, $p < .001$) and for the total sample, 1.8 – 2.3 ($df = 128$, $p< .001$). Thus, there was an indication of disagreement to statement, but less than before. The results revealed also differences between genders: the females were more critical in their evaluation than the male and their reflections were directed toward many different aspects of the form and content of program and teaching. Obviously, it is an indication of the female students' background, and their greater sensitivity to the environmental treatments and time effects in the contextual frame of the Faculty and outside it. These findings suggest that gender and contextual factors were associated with student's program evaluation.

14.4.2 Classification of students into their contextual combined course groups (1) and (2) on the basis of program evaluation variables scores

In the second stage, a discriminant analysis was undertaken. 286 cases were processed, and two cases hade at least one missing discriminant variable (in the male population). These analyses (a) revealed the dimensions on which curriculum group 1. and 2. differed, (b) identified the variables that most contributed to differences as dimensions, and (c) revealed the extent to which group membership could be classified by using students' course evaluation variables. The table in Appendix 10 reports the significant discriminant functions derived from rating variables for the male (Wilks' Lambda = .25, df 33, $p < .000$), female (Wilks' Lambda = .412, df 33, $p< .000$) and for the total sample (Wilks' Lambda = .62, df 27, $p < .000$) and the standardized discriminant coefficients in functions for the variables in the order of selection step entered to

function. From the 58 students' rating variables used in this investigation, step-wise analyses selected for the male population 33, for the female 33 variables and for the total population 27 variables.

These variables formed significant discriminant functions at the one percent level, which distinguished between subjects of the two combined course populations. The canonical correlations of functions (Rc) were in the male group .87, in the female group .76 and in the total sample .61. In each setting they accounted for 100 percent of the explained variance in each setting.

The test of equality of group covariance matrices was conducted using of Barlett's BoX'sM and it was for the *male* sample, 1049,8, F=1.24, p< .0001, df 561, 28526.9; for the *female* sample 826.46, F=1.14, p< .01, df =561, 54734.1; for the *total* population 515.16, F=1.22, p< .00, df 378, 166304.6. Table 61 displays the group means of the canonical discriminant functions.

TABLE 61 The canonical discriminant functions evaluated at group means among three data sets

Course group	Function male	Function female	Function total
(1) Before study reform	-1.44	-0.91	0.62
(2) After study reform	2.09	-1.51	-0.97

Table 62 summarizes the classifications based on discriminant function analyses for the male, female and total population. As found on the basis of student course rating variables, the analyses succeeded well in classifying 96 percent of the male, 91 percent of the female and 76 percent of the total sample into their in-group or out-group contextual combined groups (1) before and (2) after the study degree reform. The sensitivity or classification power of the male and female discriminating functions was as strong. For the females, it was identified to be stronger in the later sample (2. 1980's) classification as in the first (1970's) (89 % - 94 %). This was also a clear indication of a good level validity of student program evaluation and of the measuring instrument used in this investigation as well as a means of program intervention.

An overview of these analyses and findings revealed a significant overall contextual effect on students' ratings of instruction and also differences between male and female reflection toward teaching. Comparison of classifications by gender based on ratings also revealed a great variability in variables selected to discriminant functions: only 45 percent of the variables were the same in male and female functions and only 37.4 percent of these were selected to the function of the total population. *However, these findings did not indicate whether the combined group student ratings differed from each other in the course implementation dimensions, nor what was the relationship between the contextual factors, sex and course group.* Therefore, further investigations were needed for the determination of variance in these criterion variables.

TABLE 62 Classification power percent of grouped cases (1) and (2) before and after the study reform correctly classified by using students microteaching course evaluation variables; discriminant analyses for the male, female and total population groups

Actual group	Number of cases	Predicted group membership	
Male *		(1)	(2)
Group 1	67	65	2
Years 76,78,80		97%	3%
Group 2	46	2	44
Years 81,82,88		4%	96%
Percent of grouped cases correctly classified:			**97%**
Female		(1)	(2)
Group 1	106	94	12
76,78,80		89%	11%
Group 2	64	4	60
Years 81,82,88		6%	94%
Percent of grouped cases correctly classified:			**91%**
Total **		(1)	(2)
Group 1	173	113	46
Years 76,78,80		77%	23%
Group 2	111	26	85
Years 81,82,88		23%	77%
Percent of grouped cases correctly classified:			**77%**

*) 3 cases had at least one missing discriminating variable. 113 cases were used for printed output.

**) 283 cases were used for printed output. *) 2 cases had at least one missing discriminant variable.

14.4.3 Factor structure

In order to reduce the number of students' rating variables and to identify more general program "implementation" dimensions, the ratings of the sample of four intake course populations (1974/1976, 1976/1979, 1979/1982, 1986/1988) were subjected to Principal Axis Factor Analysis and oblique rotation. The results are presented in Appendix 10.4. The correlation matrix of the 65 items was examined using the determinant coefficient ($p = .0005$). 14 items were deleted based on the criterion of intercorrelations ($p < 0.05$) and communality (>.20). The determinant of correlation matrix was 0.000, and Barlett's test of homogeneity of the variance was used as a further check. 209 valid cases were processed, with 4.3 percent missing cases (n=203). Based on eigenvalues and interpretation of factor structure, a seven-factor model for the rating scale was accepted. The factors accounted for 41 percent of the variance. *Factor scale reliabilities* were computed for internal consistency using Cronbach's alpha coefficients. These were for the first factor .92 (18 items), for the second .86 (11 items), for the third .84 (9 items), for the fourth .73 (5 items), for the fifth .71 (4 items), for the sixth .70 (2 items) and for the seventh .50 (2 items). Generally, the internal consistency of factors was good. The set of tables presented in Appendix 10 contains means, standard deviations and two-tailed t-tests between the male, female and total populations and the list of rotated loadings

of the seven factors. Only few individual loadings and their means and standard deviations will be presented here to illustrate the structure of factor space.

The first general factor accounted for 55.5 percent of the common variance, and 18 items were loaded on this bi-polar factor dimension. It was labelled *Course in curriculum program* (+) – (-). The two items with highest loadings on the positive and negative pole best represent this dimension and no. 35 reads: "The course as such is rather useful" .66 (M = 4.0, SD = 1.0) and no. 30 "The whole course is useless in education of P.E. teachers -.67 (M = 1.9, SD = 1.0). Items connected to student motivation, were also included: item no. 9 "Exercise tasks have been sensible .63 (M = 3.2, SD = 1.1), item no. 39 "It was easy to kept interest in the subjects during exercises presented at lectures" .57 (M = 3.5, SD = 1.2) and item no. 46 "Lecture course give me new ideas the least about P.E. teaching" .50 (M = 3.5, SD = 1.1).

The second bipolar factor was connected with quality of teaching and students' own experiments of goal-directiveness. It was labelled *Clarity of presentation* (+) – (-). The best items which had high loadings and were almost "pure" were (item 3) "I was able to get right idea of objectives of lecture course from the beginning" .68 (M = 4.45, SD = 1.1) and item 4, "I was aware of objectives of the exercises from the beginning .60 (M = 3.9, SD = 1.0).

The third bipolar factor was connected to the structure of the study unit. It was labelled *Theory-practice integration.* The best item, which loaded on the positive pole of this dimension, no. 56 states: "Exercises clarified the theoretical issues" .51 (M = 3.9, SD = 1.0) and on the negative pole item no. 37 reads: "Lectures and exercises were integrated well" -.50 (M = 3.2, SD = 1.1).

The fourth factor was connected with program intervention strategy issues and it was called *Structural outline of teaching episodes and feedback* (+) – (-). It loaded with four added specific items connected to the program, including item 65 which states: "The structural outline facilated the construction of the plan for teaching .66 (M = 3.8, SD = 1.0) and item 61, "The task of evaluating teaching was useful .47 (M = 2.4, SD = 1.0). The fifth factor was pure, independent of other items and it was connected with contextual issues. It was called *Time reservation for events* too short" (+) – (-). The best item loaded on this dimension was no 17, which states: "Exercises proceeded too quickly" .56 (M = 2.7, SD = 1.2), and the second item no 18 which reads: "Too little time was spent on analysis of feedback" .53 (M 2.8, SD = 1.3). Also the sixth factor dimension was pure and independent of other items, and it was connected with the instructional material. It was labelled *Handouts in lectures, usefulness* (+) – (-)". The heavily loaded item no 28 reads: "Handouts outlining the content and lectures were useful from the point of view of attaining objectives of lectures" .78 (M = 4.3, SD = 1.0). Also the seventh factor was almost pure and loaded with two items also connected to the instructional material issues in lectures. It was called *Use of AV-material in theory-practice integration* (+) – (-)". Item no 19 states: "Lectures should have included more audiovisual equipment" .42 (M = 2.9, SD = 1.3) and item no 26 "I was generally bored during lectures" .58 (M = 3.3, SD = 1.1).

14.4.4 The variance in factor scores by gender and course group

In the first stage, in order to determine which course groups differed from each other, the multiple range test (Schéffe, p< .05) and one-way analyses of variance were conducted. Results in Table 63 indicate how the four-course group male and female students evaluation differed from each other on seven factor dimensions. Firstly, it can be noted that in regard of the total population, there were no significant differences between the course groups in the first general dimension (F1) "course in curriculum program". However, in gender course group comparisons, differences were found between course groups in all seven dimensions among the females in ten comparisons and among the males in six, and they were mostly in different dimensions.

TABLE 63 The comparison of students' program evaluation factor variance across the four course and gender course groups (two-way ANOVAs) and Schéffe test, n=203

Students'		1976		1979		1982		1988		ANOVA Source of variance: Schéffe test*) (p<.05)				
										Sex	Course Group	Inter-action		
course rating		M	F	M	F	M	F	M	F	df=1	df=3	df=3		
factors		n=28	n=45	n=22	n=34	n=10	n=19	n=21	n=21	F	F	F	M	F
F1 Course in curriculum														
program (+) – (-)	M	-.01	.18	-.01	.16	.11	-.61*	-.32	.20	0.94	2.05	2.50		76-82
	SD	.78	.91	.72	1.00	.46	1.31	1.01	.84					
F2 Clarity of goal													76-88	
precentation (+) – (-)	M	.56	-.16***	-.08	-.23	.96	-.12**	-.19	-.27	14.53***	3.83*	2.98*	79-82	
	SD	.91	.85	.79	1.02	.61	.53	1.03	.57				82-88	
F3 Theory – practice														
integration (+) – (-)	M	.10	.02	.43	0.3	-.58	.37***	-.26	-.26	0.11	2.01	4.06**	79-82	
	SD	.95	.80	1.02	.91	.42	.88	.90	.78					
F4 Structural outline for														82-88
teaching episodes and														76-82
feedback (+) – (-)	M	.28	.26	-.01	-.34	.29	-.62**	-.17	.18	1.58	5.13**	3.07**		76-79
	SD	.74	.72	1.04	.95	.47	.88	.76	.66					
F5 Time reservation														79-82
for events	M	.32	-.06	-.44	-.51	-.01	.40	.31	.52	.19	11.83***	2.03	76-79	79-88
too short (+) – (-)	SD	.97	.52	.74	.82	.60	.93	1.03	.74					76-88
F6 Handout in lectures,														
usefulness (+) – (-)	M	-.14	.23	.27	.39	-.23	-.05	-.56	-.36	3.69	7.31***	0.25		76-88
	SD	1.04	.56	.77	.62	.74	1.16	1.09	.82					79-88
F7 AV-material in														
theory-practice														
integrations	M	.20	-.46***	.25	-.11	-.42	.05	.47	.28	7.84**	4.58**	3.54*	82-88	76-88
(+) – (-)	SD	.82	.75	.73	.78	.70	.80	.78	.81					

Values are means, SDs, Fisher's F-statistic (variance ratio); *,**,*** p<0.05, 0.01 and 0.001 respectively
*) F-values were calculated after applying Barlett's test for homogeneity of variance Schéffe (1967)

The male groups' ratings differed most frequently in pair-comparisons connected to the quality of teaching, such as "Clarify of goal presentation" (F2), whereas among the female course groups to more specific factors such as "structural outline for teaching episode and feedback" (F4) and "time reservation for events" (F5) as well as to usefulness of handout in lectures (F6), and AV-material in theory-practice integration (F7). The extreme course group (1976 and 1988) students' rating differences appeared most frequently in this comparison. The results, presented in table 63, indicate also how variance among the seven-factor dimension varied by sex and course group.

14.4.5 The relationship between student program evaluation and contextual background – sex and course group variables

Table 64 reports results of the two-way ANOVA factor scores by sex and combined course group (1) before and (2) after the study degree program reform and comparison by gender. As the table shows, there was only one dimension without such contextual effects on rating differences - factor four - which was connected to the basic elements of the microteaching course, labelled *"Structural out-line for teaching episodes and feedback"*.

TABLE 64 The comparison of students' program evaluation rating factor variance among students' grouped by decade 1970's and 1980's course and gender course groups (two-way ANOVAs) and two-tailed t-test

Varimax factor		(1)Decade 1970's (n=129)		(2)Decade 1980's		ANOVA: Source of variance:				
		M (n=50)	F (n=79)	M (n=31)	F (n=40)	Sex df=1 F	Decade course df=1 F	Inter-action df=1 t	M-F (1) df=127 t	(2) df=69
F1: Course in curriculum program (+) – (-)	M	-.01	.17	-.18	.18	.69	3.96*	.42	-1.12	0.01
	SD	.74	.95	.89	1.15					
F2: Clarity of goal presentation (+) – (-)	M	.28	-.19	.18	-.20	11.71***	.13	.13	2.83**	1.93
	SD	.92	.92	1.06	.55					
F3: Theory – practice integration (+) – (-)	M	.25	-.02	-.36	.04*	.04	3.08	6.78**	1.64	-2.18*
	SD	.93	.85	.61	.88					
F4: Structural outline for teaching episodes and feedback (+) – (-)	M	.15	.00	-.02	-.20	1.71	2.28	.02	0.94	0.94
	SD	.89	.87	.71	.87					
F5: Time reservation for events too short (+) – (-)	M	.02	-.25	.21	.46***	.22	17.41***	3.90*	1.60	-1.24
	SD	.94	.71	.91	.83					
F6: Handouts in lectures usefulness: (+) – (-)	M	.04	.30	-.45*	-.21	4.18*	16.37***	0.01	-1.89	-0.99
	SD	.95	.59	.99	1.00					
F7: Use of AV material in theory-practice integration (+) – (-)	M	.22	-.31	.18	.18**	8.81**	5.20*	4.86*	3.78***	0.03
	SD	.78	.78	.85	.80					

Values are means (SD), F-Fisher's F-statistic (variance ratio) variance are not equal between groups; *, **, *** p<0.05, 0.01 and 0.001 respectively
M=male
F=female

It should be noted that these features related to content, methodology and program intervention strategy were not changed in the contextual reform. This was an indication of the stability and validity of students' program evaluation. Moreover, the only factor dimension where the source of significant factor variance was the curriculum group was the first general factor. This factor, *"Course in curriculum program"*, indicates the relevance of the study unit in the frame of the faculty: ($F(1,199) = 3.96$, $p < .05$). This indicates clearly that there was relationship between students' program evaluation and changes made in the program contextual setting. The only factor dimension where the sex of students was the only source of significant variance was factor two, which was connected to the quality of teaching, *"Clarify of goal presentation"* ($F(1,199) = 11.71$, $p < .001$), and it was experienced to be different (better than before) in the

male students' ratings. As significant difference between male and female ratings was identified in this dimension ($F(127)=2.83$, $p < .01$). The strongest course group effect was found in factor dimension five *"Time reservation for events"* ($F(1,199)=17.41$, $p < .001$) with interaction by sex ($F(1,199) = 3.90$, $p < .05$) and it was experienced to be too short especially by the female students. Factor six was connected to problems of *"handouts in lectures, usefulness"*, and the source of variance in this dimension was course group ($F(1,199) = 16.37$, $p< .001$) and sex ($F(1,199) = 4.13$, $p < .05$). The experiences of usefulness of handouts were different (diminished) especially in the male students' ratings. In factor seven, concerning *use of AV-material in theory-practice integration* the sources of variance were sex ($F(1,199) = 8.81$, $p < .01$), course group ($F(1,199) = 4.86$, $p< .05$) and interaction of the first two ($F(2,199) = 4.86$, $p< .05$). The females found the need of AV material to be more urgent than the male students ($F(1,127) = 3.78$, $p < .001$).

14.5 Summary, conclusions and discussion of pilot study III B

Summarizing the main findings of the multidimensional analyses: firstly, the assumption of the contextual effects on students' program evaluation was supported in the results of this particular sub-study. The changes made in the curriculum program of the Faculty reflected in changes in different program criterion evaluation dimensions. The ratings based on students' own experiences concerning *the relevance of the study unit program in the faculty curriculum differed* ($p < .05$), indicating a decreasing trend after the curriculum reform. Overlapping was experienced within revised the study program more than before. Obviously, this is, in fact, an indication of the study unit program's congruence with the revised faculty curriculum program. Moreover, when the organization of study units in the faculty curriculum plan was revised, and e.g. time reservation for microteaching practice events was diminished by 30%, this reflected clearly also in students' ratings. The effect of the contextual course group was strongest in the evaluation dimension *"Time reservation for events too short"* (F5) ($F = (1,199) = 17.41 < .001$). It was experienced as a great problem for students' already before the study degree program reform and more so after it. The results of this study confirmed the findings from another curriculum evaluation study conducted after the study degree program reform at the faculty (Rantakari & Tiainen 1983).

Moreover, the contextual factor, including students' background, reflected strongly on student program evaluation. The assumption of rating differences between males and females was supported in this study. Many interaction effects recognized in the variance in evaluation dimensions were an indication of gender and/or subpopulation differences. The student selection procedure, e.g. stage 1. sum scores (school success), produced different gender populations. Obviously the females with a higher level of cognitive capacity were more sensitive and critical to the environmental setting where the preservice course was conducted than the male students. It was clearly

identified in the first stage analyses. Furthermore, the results of discriminant analyses also revealed clear gender differences: 95% of the male and 91% of the female students, but only 77% of the total sample subjects could be correctly classified into their own criterion groups (in-group or out-group) based on student program evaluation variable scores (1) before and (2) after the study degree program reform at the faculty. This result is also an indication of the good construct validity of the measuring instrument, the questionnaire, as a means of intervention in the study unit program. The results concerning rating differences by gender and curriculum group confirmed the findings from the earlier presented study Phase III A (see also Heinilä 1988, 1992a), and the results obtained by Telama et al. (1988). Also the findings paralled with results obtained in program evaluation studies, connected to the assessment of differences between persons reflections to the environmental treatments in teacher training programs (Hanke 1980b, Siedentop 1981).

Additionally, it might be noted that students of both curriculum groups were satisfied with the intentionality and usefulness of the course in the study program (item 35, M = 4.0, n = 283). The organization and coordination of lectures and demonstrations was observed to be good (item 56, X (283) 3.9) and they thought that during this course they had become aware of errors and weakness in their teaching behaviour (item 59, M = 3.8, n = 283).

Although the student program evaluation was valid and the questionnaire used appeared to be an appropriate scale for assessing rating differences within the study unit program contextual setting, further development of the scale examining the realization of faculty curriculum study units in terms of context, target behavior, content process and time reservation frame would be necessary (see e.g. Rantakari & Tiainen 1983, Silvennoinen et al. 1991).

Nevertheless, given the complexity of attitude and study motivation formation, it is recognized that other factors beyond those presented in this program evaluation study, such as students prior experiences at school and in the social setting of the teacher education program (see e.g. Hendry 1969, 1978, Laakso 1975, Martens 1987, Martin et al. 2001, Silvennoinen et al. 1991, Telama et al. 1988, Whitehead 1980) may also have been important contributors toward values and study motivation formation, which reflected also on students' own experiences and their program evaluation. Further research would be needed to enlarge the perspective in relation to other aspects of students' study experiences with teaching behavior development. Such understanding might be informative from the teachers', students' and administration's perspective in planning future curriculum programs (see e.g. Silverman 1991, 235 – 236).

15 SUMMARY AND CONCLUSIONS

15.1 Overview

In this chapter the main results of the dissertation are summarized and some conclusions are drawn. The summary first recapitulates the main findings on the primary research problems. Then some of the strengths and limitations of the study are critically discussed. This is followed by an outline of areas suggested for further investigation. Finally, some possible implications of the study for research on teacher education and on the teaching of physical education are discussed.

In January 1974 the Department of Physical Education of the University of Jyväskylä introduced, on an experimental basis, a new type of practice teaching using procedures of didactic observation and microteaching based on "Human Interaction Model". A new preservice course emerged from part of degree requirements and was intented to be given during the last term of the third year as an obligatory course (45 hrs) (Heinilä 1977b) after the course of didactic observation course (30 hrs), and after the study reform (1978-) these courses were combined (2 study weeks). It was taught by the author, a Faculty member, until the year 1991. Thus, this dissertation covers a long-term research project focused on obtaining information for program development, for its control and implementation in a social setting so that the intented goals and competences could be optimally attained and theory and practice brought closer together (Telama & Vuolle 1976, Telama et al. 1980). The "didactically thinking" P.E. teacher was thus, already in the late 1970s, the "ideal teacher" according to the expectations presented in the study plan, as well as in the new teacher-training program. This kind of teacher more called at recently "a reflective teacher'. She/he will not only answer right to the question "How", but asks "What" and "Why" – in response to variation of instructional situations. (see Hupé 1995)

The development of educational programs is a multistage process at several levels and should be based on scientific research. This dissertation comprised two sections: the first section concentrated on meta-level questions on concepts, models, and methods for the development of an Interaction Analysis

System for physical education research and teacher education and the second section on the application of the System to teacher education and on a longitudinal long-term program evaluation in contextual variation. The research project was supported by the Ministry of Education, the Finnish Cultural Foundation, and the Ellen and Artturi Nyyssönen Foundation, and The University of Jyväskylä.

The main purpose in the first phase of this study was to develop and test a system for describing instructional procedures in physical education. Its aim was to construct a method for providing good descriptions of teacher-student interactions in P.E. classes, rather than to test theoretical hypotheses or evaluate the effects of such interactions.

Thus the study had a clear methodological orientation. Drawing mainly on interaction theories of the teaching-learning process and on available research, it sought to develop a theoretically justifiable system for describing and analyzing what happens in the physical education classroom. The second research task was to critically test the reliability and validity of the constructed system. The third research task was to develop a paradigm for describing the activity form and the formal proprieties of the instructional process in physical education.

In the second phase of the study the first research task was to develop a teacher-training program based on the theoretical framework and to study the application of this system to teacher education through curriculum evaluation in microteaching. The second research task was (1) to control the intervention strategy and the basic elements of the program, accountability of the modified observation instrument PEIAC/LH-75 II, and study its construct validity and sensitivity as a feedback instrument in connection with learning non-directive teaching skills, (2) to study the applicability the entry teaching behavior rating scale, (3) to study the quality of students' "ideal" P.E. Teacher characteristics rating scale and (4) to study the functioning of the questionnaire for students' program evaluation. In the final phase, the purpose was assessment of the predictive validity of the study unit program conducted (1974 – 1991) in contextual variation, before and after study degree program reform at the Faculty (1978), combined with student program evaluation.

The approach used in this study was primarily based on the theoretical and practical work done by Flanders (1965, 1970) with reference to his paradigm and the research literature related to the original FIAC system and its several adaptations. The impetus for the present study came from the DPA Helsinki project. Professor Matti Koskenniemi encouraged the author to start an enquiry into interaction in the gymnasium. Professor Erkki Komulainen's exhaustive and perceptive methodological studies on classroom observation served as a model whose sophistication is worthy of emulation but not easy to achieve. The author's experience of the course of didactic observation at the University of Helsinki 1969 was a turn point. In agreement with Koskenniemi, the development of 'didactical thinking' is obviously an important educational aim. According to him, a didactically thinking teacher works like a researcher. Thus, the author learned "reflective teaching" by using the concept in the 1990s (see Paré 1995) in this research project.

Section I

15.1.1 Developing an interaction analysis system for physical education classes

In the first section, a study of related research literature and consideration of the specific character of physical education indicated a clear need to adapt the Flanders Interaction Analysis System based on "Human Interaction model", the and observation instrument FIAC. Since movement is an integral part of the instructional processes in P.E. classes, it was obviously necessary to be able to take into account how movement communicates and influences; also the social form of instructional process in P.E. was found to be one of the main issues (Heinilä 1971). Consequently, three clusters were included in the developed PEIAC/LH-75 system (Heinilä 1977a). The first cluster describes teacher and student talk and teacher's silent activity. The second cluster deals with students' collective movement activity/passivity and social access. The third cluster records the social form of the class. These three clusters contain 12, 8, and 7 categories respectively, altogether 27 categories. Since this cluster arrangement required triple coding, a six-second interval was used instead of the three-second interval employed in the FIAC system. The decision was based on the consideration that three seconds was too short a time for the complex coding required of coders and not needed for the observation of two other clusters. The PEIAC/LH-75 analyses system provides primary and secondary information through a specially adapted computer program.

The data was collected in such a way that the developed system could be tested in a number of ways. The data used to evaluate the descriptive adequacy of the developed observation schedule and observation training consisted of 24 P.E. lessons, altogether 28,800 six-second time units. The objectivity of coding was assessed by studying the level of agreement between six trained (20 h) outside observers. The sensitivity of the system to faithfully reflect similarities and differences in P.E. classes was studied by including four different areas of subject matter (gymnastics, apparatus, rhythmic movement-expression, and ball games) in the 24 lessons. For the same reason, boys' and girls' lessons from three different grade levels (lower grades: 1 - 3; middle grades: 4 - 6; and upper grades: 7 - 9) were sampled. The construct validity of the system was studied by examining the patterns of data obtained through primary and secondary analyses in the light of the posited model.

15.1.2 The reliability of PEIAC/LH-75

The first aspect of the reliability of the developed system dealt with the objectivity of coding. It was studied in both live and videotaped situations. The results indicated that the inter-coder agreement was somewhat higher with the videotaped material than in the live situation. This might be explained by the fact that the situational complexity was reduced in a videotape recording.

The second aspect of reliability dealt with the objectivity of coding in terms of inter-coder agreement, within-coder constancy and between-coder constancy. The method used was Scott's *pi* coefficient. Summarizing the main results, the average level of mean coefficient values was rather low and varied according to cluster: Cluster I, .61; Cluster II, .65; and Cluster III, .69. The inter-coder agreement was .65, within-coder constancy .69, and between-coder constancy .60 when the two observations of the videotape recordings (T_2 and T_3 were compared.

The third aspect of reliability focused on reliabilities of the various individual categories, operationalized as inter-coder agreement, and assessed by means of Kendall's coefficient of concordance (*W*). This analysis indicated that agreement was fairly high, with 23 out of 27 categories yielding a value of *W* significant at the .01 level (Chi Square test). In all coding situations, however, the coefficients of four categories of infrequent occurrence (I/03), and confused situation (I/12, II/8, and III/7) were not statistically significant.

As a fourth aspect, the variability of coder's or "validity of coders" was studied using discriminant analysis. The first two of the five discriminant functions were statistically highly significant and a third one nearly significant (58%, 21%, and 11% of total discrimination, respectively). The first discriminant function distinguished those observers who made a wide use of the categories of verbal communication from those who used only some of these categories. The second function separated coders by their coding choice in a situation, which might be variably interpreted as either confused or as displaying spontaneous student activity with teacher's silent participation. The third discriminant function distinguished coders who described a sequence of verbal and nonverbal communication by using also infrequently occurring categories from those who employed only frequently occurring categories.

The results indicated that there may be an inverse relation between reliability and validity in the case of observation research. Crude coding may be advantageous in terms of reliability, but be detrimental to coder validity. Observation is a skill and its learning might be related to personal characteristics, knowledge and expectations of the observer as well as to the factors of teaching learning conditions.

It was concluded that the three-dimensional measuring instrument (PEIAC/LH-75) was reliable when estimated by using a nonparametric coefficient of concordance, *W*.

15.1.3 The validity of PEIAC/LH-75

The first aspect of the validity study of PEIAC/LH-75 addressed the question of construct validity. To stress this crucial aspect of all research, a model was developed to define the overall research strategy for the study. This model served as a guide (1) in specifying the entry situation by defining a theoretical and conceptual framework, (2) in constructing a set of exhaustive and mutually exclusive observable behavior categories on the basis of the conceptual framework, (3) in selecting the unit of observation and in developing an adequate coding procedure for accurate use of the system, (4) in selecting the

unit of analysis. The instrument was developed on the basis of a detailed review and analysis of available literature on research on classroom interaction. This critical survey showed that the Flanders one-dimensional verbally oriented system needed to be complemented. The feasibility of a multi-dimensional coding system was affirmed in pilot work (Heinilä 1970, 1971, 1974).

Construct validity is often determined in an indirect way. The researcher uses a theory to establish a set of hypotheses about how the data should behave. For instance, the researcher predicts certain internal relationships between measured variables: high, intermediate or low correlations. A construct-valid instrument will produce scores that correlate only with those variables with which, on the basis of theory, it should correlate, and the scores of those variables to which it should not be related will not correlate with it (convergent vs. discriminant validity). Similarly, a construct-valid instrument should distinguish between groups that are known to behave differently on the construct under study.

In the primary analyses, it was noted that all of the PEIAC/LH-75 categories were used in coding. Thus, the instrument did not appear to contain superfluous categories. Second, 22 statistically significant differences out of the total of 27 categories were found as functions of frame factors: 4 between the two teachers of the sample, 5 between grade levels, and 13 between the various subject areas of physical education classes. Third, matrix analysis showed the interaction sequences to be different in the three clusters, as expected, providing a good description and yielding more information concerning critical teaching behavior. In the first cluster, more than half of all sequence pairs were in the steady state cells while the corresponding figures were more than 80% and more than 90% for Clusters II and III, respectively. This indicates that decisions concerning social form, division of labor and responsibility as well as the forms of students' collective activity/passivity were the general dominating features of teacher behavior.

As another indirect indicator of construct validity, teacher directiveness decreased as a function of grade level while teacher's silent guidance, participation, use of student ideas, and pupil responsibility increased. Also, the variety of critical sequence patterns increased and was strongly related to the content area of physical education.

In the secondary analyses, 18 indices were computed to reduce the primary descriptive analyses. These indices were based on unit coding and the statistical procedures were based on category frequencies, percentages, and ratios. They were computed separately from the matrices of the three clusters. The results indicated that in all 18 parameters of PEIAC/LH-75, statistically significant differences (Mann-Whitney U-test) were found as a function of the key frame factors: teacher (5 statistically significant differences), grade level (6), and subject area of physical education (14).

Factor analysis yielded seven factors, which accounted for 68.6% of the total variance. The variables in the factors were concerned with the following: Factor I, with indirect nonverbal integrative idea generation; Factor II, with the intensity of the teacher's verbal direct guidance; Factor III, with the uniformity of the teacher's nonverbal guidance as opposed to the specificity of verbal

supportive supervision; Factor IV, with the direction of teacher-pupil communication; Factor V, with spontaneous student activity; Factor VI, with subject-centricity vs. process centricity; and Factor VII, with teacher's response behavior focused on individuals vs. groups.

Grouping analysis was used to relate lessons to the extracted factor dimensions. This made it possible to establish the type of lesson that was most characteristic of each factorial dimension. Through this procedure, empirical knowledge of what the lessons were like was obtained. Six structurally homogeneous lesson groups were formed and compared. The main elements of the goal directed interaction process was represented by the way of learning (cognitive, affective and psychomotor proprieties of verbal and nonverbal communication). It was confirmed that the various subject areas in P.E. and/or the teacher instructing them reflected strongly in regulating interaction. The lessons with closed/open subject area differed clearly.

The starting point for a further exploration of the predictive power of the categories of three clusters was estimated by using discriminant analysis, based on the means and dispersions of the grouped lessons (n=144). Five discriminant functions were extracted: DFI, range of ideas for students (closed vs. open); DFII, level of structuring (high vs. low); DFIII, level of intensity of guidance (high vs. low); DFIV, level of specificity of non-directive guidance (high vs. low); and DFV, media of non-directive communication (nonverbal vs. verbal and attributing of teacher's response to individuals as opposed to groups). The analysis selected 16 out of the total of 27 categories and set them in sequence according to how much they increased the model's discrimination power. The categories of the second cluster (students' collective activity/passivity and social access, and the categories of the third cluster, social form) proved to possess the highest discrimination power.

Through an extensive set of explorations, summarized briefly in the above, it was concluded that (1) the instrument possesses a definite degree of construct validity, and that (2) it is sufficiently sensitive to discriminate aspects of direct-nondirect teaching behavior in physical education classes.

15.1.4 Activity forms in the paradigm of PEIAC/LH-75

On the basis of investigations for the development and validation of Interaction Analysis System, PEIAC/LH-75, a paradigm has emerged in which intentionality and social structure are considered the main elements of the instructional process in physical education classes. They are represented by the content and manner of learning (cognitive, affective and psychomotor proprieties of verbal and non-verbal communication). The social structure manifests itself as the teacher and student roles that regulate the interaction. These roles are reflected in the division of labor and responsibility between teacher and students and in the grouping of students. The other characteristics of the PEIAC/LH-75 based on the theory of Flanders are the social emotional climate of the gymnasium, and students' cognitive, affective and psychomotor engagement, teacher's authority in use and flexibility shown by the teacher in striving at educational aims, and objectives as prescribed in the curriculum

plan. Also the media of communication is an important aspect of the instructional process in physical education.

The model of "Activity forms in the paradigm of PEIAC/LH-75" was used as the frame of reference for the organization of instructional process by using the system to the P.E. teacher-training program study unit at the Department of Physical Education of the University of Jyväskylä.

Section II

15.2 The application of PEIAC/LH-75 to teacher education and program evaluation

The second section of this research program comprised three successive phases: the first research task was to apply the instrument that had been developed to the task of training the future teachers of physical education. This was carried out through a curriculum and intervention strategy model for a preservice teacher training study unit, the course of didactic observation and microteaching. A scientific management of the teaching process was set as a goal of the new system of the training of P.E. teachers. Research had indicated that the systems of interaction analysis as tools in teacher education offered better opportunities of achieving this goal, the interaction of theory and practice.

15.2.1 Pilot study I: curriculum evaluation

The development of new programs of practice teaching presupposes the controlling and evaluation of their basic elements. The purpose of the first phase of this study was to evaluate and compare two curricula whose purpose was to develop the verbal indirect teaching behavior of student teachers. The congruence between intended and actually occurring outcomes was analysed, described and judged in terms of process criteria. The curricula of the courses differed in terms of the following elements: (I) information about (models of) target behaviour (written, audiovisual), (II) timing of instruction of theoretical considerations (before/during the course), (III) size of training groups (5 - 10), (IV) length of microlessons (5 - 10 min.), and (V) number of microlessons (2 - 3).

The data covered the courses of microteaching arranged by the faculty in 1974 and 1976 and the subjects were male and female students who started their studies in 1971 (n = 48) and in 1974 (n = 74), altogether 275 microlessons.

The measuring instrument (PEIAC/LH-75 II, Heinilä 1977b) had been constructed for teaching and testing purposes and it was used in a somewhat modified form. It was derived from Flanders' FIAC-system and contains two clusters, speech and movement, and altogether (16+2) 18 categories. A double coding was made at six-second intervals. It made it possible to give information about target behavior, to operationalize model behaviour and to analyze

TV-feedback using a systematic observation method. Reliability (.78) was estimated by means of Scott's pi-coefficient. The category frequencies, indices and student evaluations of courses by using a questionnaire were compared using analysis of variance, t-test (ANOVA), and Chi Square test. The reliability of the questionnaire in terms of homogeneity (Cronbah's alpha) was computed on the basis of item-test correlation for students in four populations and it varied between .56 and .92. (cf. Heinilä 1988)

Statistical comparisons of the outcomes based on process criteria, of each course showed clearly that the revised course program differed from the first version on the level of realization. The success of the program was reflected in (a) a decrease of teacher talk, (b) and increase of teachers' silent didactic activities, (c) an increase in teacher response behavior, and (d) a decrease in content emphasis. The increase of indirect behavior was evident in the second session, in which the teachers offered the pupils more opportunities to create ideas and solve problems, observed pupil responses, and took advantage of these responses in the progress of the topic treatment.

The students of both sessions were asked to evaluate the course. A comparison of the responses indicated that, although the students in both sessions were generally pleased with the content, timing and organization of the course, the second group clearly benefited from the revisions that had been made. They felt that the course had opened a new outlook and that they had learned to discriminate between teaching patterns in observing and coding feedback. They thought that the course had been useful, making them aware of errors and weaknesses in their teaching behavior. Most importantly, they reported that they intented to use the teaching patterns they had learned in their future practical teaching.

15.2.2 Validation of the basic elements of the microteaching program (II)

The objective of this particular study was to assess the revised program intervention strategy model's basic elements A-D: how well the modified observation instrument PEIAC/LH-75 II functioned in relation to the entry teaching rating scale and the rating scales concerning students' expectation on the "ideal" P.E. teacher characteristics.

(1) Pilot study II A: the validation of an observation system: a multivariate approach

In the first phase, an attempt was made to determine the descriptive adequacy of the category system, PEIAC/LH-75 II, by using the six models constructed and operationalized as target behavior in microteaching. The subjects were students (n=74) of the revised microteaching course. The multiple discriminant analysis technique was used for the evaluation of the scores obtained from the observation of six different model groups in the first and second videotaped 10 min microlesson (n=148) by trained observer (Scott's Pi .78). In this phase, four discriminant functions separating the criterion groups "known" to behave differently were found. The discrimination of the three first could be considered

highly significant. The share of total discrimination for each discrimination function was: 47.7%, 27.5%, 16.9%, and 5.7%. The program selected 9 of the 18 categories of the modified instrument and set them in sequence according to how much they increased the model's discrimination power. It was also noted that four of the added five subscripted categories were selected to the model.

Most of teaching models could be placed on dimensions formed by the discriminant functions reflecting their specific aspects of indirective teaching. The hypothesized correspondences of the categories to the 6 teaching models were shown to be successful for models 2, 3, 4 and 5. In teaching models 1 and 6, the category-reality correspondence was not so clear in this data. The structure of the discriminative model was related to the structure of the measuring instrument (Rc values .70, .60, .50 and .32) and produced a clear sequence predicting grouping of microlessons in accordance with different "known" models used in this study (cf. also Heinilä 1990).

(2) Pilot study II B: investigation of the construct validity of an observation instrument – a multivariate approach

In the second phase, the construct validity of the modified observation instrument PEIAC/LH-75 II was estimated in the revised microteaching setting by using Factor analysis r-technique. The subjects were students (n = 74 and n = 42), of the microteaching course arranged by the faculty in 1976 and 1988. The scores of trained observers (Scott's Pi .78), who observed 221 and 126 videotaped microlessons, (1. control, 2. and 3.) were analysed: profiles, matrices, percentages of categories, indices, correlations, factor structures by using r-technique. The factor analysis yielded three factors, which accounted for 39.7% of the total variance (18%, 13.4%, 7.4%) in the first study. The factor structure was clear. The first factor accounted for 46.9% of the variance and it was clear cut in content and consisted of variables of the first cluster. It was named *"Teacher initiation (+) vs. - teacher response behavior (-)."*Factor II consisted of categories from the two clusters. It was named as *"Channel of teacher-pupil communication: verbal (-) vs. motor (+)"*, and the third factor was named *"Teacher feedback and motivational communication (+) vs. -teacher silent guidance (-)"*. The difference of the three microlessons in relation to factor structure was analysed by using analysis of variance and t-test.

The fact that the lesson groups could be placed dimensionally (direct - nondirective teaching) and contextually (non-verbal - verbal communication) is interesting from the theoretical point of view and the cognitive orientation of the study (Flanders 1965, 1970, Heinilä 1977b). Impact on in students' teaching behavior could be verified. The results of the replicated pilot studies (Heinilä 1988) supported these findings. It was identified that the consistency of the factor structure in two different data sets was stable, and especially the power of the general factor I "Teacher Initiation (+) vs. response behavior" was high: in the first sample it accounted for 47 % in the first and for 52 % in the second data set of the common explained variance and its sensitivity to determine the variance between lesson groups was good.

The construct validity of the PEIAC/LH-75 II was also supported by the meaningful correlations of F1 scores and the ID-index of sum scores r .74 ($p < .01$) and also between students' entry teaching behavior rating scale item 3 (teacher-student interaction) scores r = .65 ($p < .01$).

Also results obtained by Akkanen (1979) in a pilot study conducted in a natural setting (n = 8) by using PEIAC/LH-75 II system and the teaching models constructed supported the results obtained in this study concerning sensibility of the instrument. Also, based on the results obtained in a pilot study conducted by Reponen (1979) it was established that (1) the order of PEIAC/LH-75 II (Heinilä 1977a, 1977b) indices revealed differences between experienced teachers with regard to the rank order of behaviors and (2) between two groups of student teachers (n=54) and between student teachers and experienced teachers. Thus these results were indicators of the good discriminant validity of the PEIAC/LH-75 II system. Simultaneously it took in to consideration pupils' collective motor engagement time (MET), which had been shown in teaching effectiveness studies to be frequently positively related with student achievement (see e.g. Borys 1986b).

It was concluded that (a) the two-cluster 18-category PEIAC/LH-75 II system for combining process and a cognitive orientation possesses a definite degree of construct validity and objectivity, and that (b) it is sufficiently sensitive to discriminate aspects of direct/nondirect leaching behavior and, to a definite degree, also aspects of nondirect teaching behavior operationalized as teaching models.

15.2.3 Pilot study II C: the teaching behavior rating scale - assessment of reliability and validity

The teaching behavior rating scale was intended to be used as a means for determing the level of students' entry teaching skills, firstly in students intake test (microteaching episode, 3.5 min) and secondly two years later before the course of microteaching, (microteaching episode, control 5 min) and also as a form of program intervention. The rating scale was used in the Faculty of Sport and Health Sciences from the year 1976 onward starting simultaneously with the course of didactic observation and microteaching (1974). The rating dimensions selected to the measuring instrument were based on research results, obtained on the relationships between teacher's personality and teaching behavior (Flanders 1965,1970, Hytönen 1973, Hytönen & Komulainen 1973, Kane 1968, Medley 1971, Rogers 1967), and also on effectiveness of teaching (Rosenshine 1971, 1976).

The teaching behavior rating scale contains four items, using a six-point scale. The rating dimensions selected were connected with following behavioral characteristics: the first item, related to teacher's "Presentation: voice quality, expression, fluency, clarity, movement behavior"; the second item, to "content presentation - Understanding of task content, phases and instruction"; the third item to "communication: teachers' interaction with pupils directiveness of main points, observation, feedback"; and the fourth item on "Creativity: originality, aptness, presentation of main points".

The reliability and validity were determined in two studies by using video recorded material (n =75 and n =42) of microteaching episodes (5 min) control before the course of microteaching and sum scores from the student selection procedure test, microteaching episode (3.5 min, n = 42). Raters were trained outside observers, post-graduate lecturers of the University.

Reliability, inter-rater agreement determined by means of Kendall's Coefficient of Concordance, W ranged for summed scores between (MD values) .75 and .68 and were statistically significant (Chi Square) at one and at five percent level in two different tests. On the item level, the reliability indices were statistically significant at one percent level (75 cases) and in the second test (42 cases) at five percent level or beyond.

The stability of ratings was determined, firstly, based on results obtained from the second observations four weeks later (n = 10, n = 12). In the first test, the reliability coefficient for summed scores, Median values, was W .96, r^2.36 (Chi Square, df=77, p<.01) and in the second test W = .84, r^2 = .56 (Chi Square, df = 9, p < .05).

Secondly, the stability of ratings based on correlations between two raters' summed scores gathered from the same population (n=42) between two years interval, and determined by means of Pearson's Coefficient of correlation and (ANOVA) two tailed t-test was good, r = .52 and statistically significant at one percent level. Furthermore, at the item level, the correlations between the two teaching episode test sum scores ranged between, r = .56 – r = .39 and were statistically significant at one percent level.

The construct validity of the ratings scale was determined by comparing results obtained from the same video record material observations by using two instruments: the rating scale variable item 3 ("teacher-pupil interaction") and PEIAC/LH-75 variables, ID-index ("teachers' response behavior") and Factor 1 scores ("teacher initiation vs. response behavior"). The correlations were r =. 42 and r = .30 and statistically significant at the one and at five percent level. The student intake teaching episode sum scores' correlation to F1 scores was .36, statistically significant at one percent level.

Based on results obtained it was judged that the reliability and validity of the rating scale was sufficiently high for further analyses and to be used as a means of student intake and interventions in the study unit course of didactic observation and microteaching. Moreover, in determining the homogeneity of the variance, of the scores in eight different subpopulations (n=205) and by comparing the results between groups, it was concluded that the rating scale was also sensitive for assessing the level of students' entry teaching skills. Also the congruence between the level of the objectives and the level of observation in the test results - obtained from two grouped sample, (1) before and (2) after the study degree reform of the Faculty (1978) - was verified. The goals the Faculty Teacher Education program and the course were congruent concerning e.g. the enhancement of students' oral delivery.

15.2.4 Phase II pilot study D: student's attitudes, "ideal" P.E., teacher expectation rating scale - reliability and construct validity

The aim of this particular program evaluation study was the validation of the measuring instrument, "ideal" P.E. teacher's characteristics expectations questionnaire. It was an adaptation of an instrument with bipolar l6 items, l -6 scale instrument developed and validated by Hytönen and Komulainen (1971) and used in an empirical study for controlling the stability of teaching style in the dimension "student-centered – teacher-centered" teaching style. It was based on ideas of Flanders (1965, 1970) and Rogers (l967). It was deemed applicable for the purpose this study, to be used as an intervention, as a means for impacting students' goal directiveness and learning gain in the course of didactic observation and microteaching, and to be used for diagnosing student's attitudes before the beginning of the training sessions. Since it was originally constructed to be used in a study conducted in the area of mathematics it was necessary to refine it. Based on the framework of the program, four items - connected to the social form, division of the labour and responsibility and teachers own motor engagement – were added (cf. Heinilä 1988, 1992a).

The reliability of the questionnaire was determined in terms of homogeneity, Cronbach's alpha, based on item test correlations for populations among four microteaching course groups (n=205). It ranged in the four factor solution from .56 to .76 on factor 1, from .45 to .49 on factor 2; from .41 to .56 on factor 3, and from .26 to .52 on factor 4. When population groups were factorized separately it ranged from .56 to .92 (cf. Heinilä 1988). This was judged to be sufficient for further analyses.

For the purpose of this study, to determine the possible multidimensionality of the scale, the material of four replicated case studies (n=205) was subjected to factorization by the principal axis method and to rotation by the orthogonal varimax technique. The four-factor pattern was chosen, and a comparison of factor variance within and between population groups was conducted by using one-way ANOVAs and two-tailed t-tests, Pearson's correlation coefficients and two-way analyses of variance. By analyzing the variance of factor scores among subpopulation and gender groups, and also among combined groups representing populations before and after the study degree program reform (1978), it was noted that there were no statistically significant differences between students' ratings in the first general factor. *Thus, the male and female students' "ideal" PE teacher personal characteristic expectations concerning the "congruence/genuineness" issue were consistent over decade and contextual course group variation.* This variable was labeled in the earlier study project "student centered-teacher centered teaching style" (Hytönen & Komulainen 1971). By contrast, in the three other factors, significant differences were noted. In factor three, the variance of factor scores was statistically significantly different between population group scores at the 0.l percent level. Thus, the "ideal" P.E. teacher of the 1980's was identified to be more "involved" than in the 1970's (the pooled variance estimate F value was -3.86, $p<0.000$). In factor two, a statistically significant difference between decade groups at the five percent level was found. The social form used by the

"ideal" P.E. teacher was in the 1980s more "class-centered", whereas in the 1970s "individual-centeredness" was dominating. It can be mentioned that the popularity of "aerobic"-system in the 1980s might be reflected in these differences identified between the ratings of two combined group female students. In the fourth factor, statistically significant course group and sex effects were found and interaction between the two (14.42 p< .000). It can be noted that "ideal" teacher's "fact-centeredness" was weighted more in the 1980s than in the 1970s, but only in the male population. (t= -2.44, p=.05).

Moreover, the results of the discriminant analysis indicated that the measurements used for determining the stability of students' entry attitudes among the male, female and total two-tailed subpopulations appeared to be adequate given the level of discriminant validity - 81%, 84% and 77%(p< .001) of the group members were correctly classified into their own combined course groups (i.e. in-group or out-group) based on their ratings on "ideal" P.E. teacher characteristics expectations. In addition, the results of these analyses and factor analyses indicated a good validity of the 20 bipolar items questionnaire (with four added items, concerning social form and movement issues).

It was concluded that these findings demonstrated the conceptual integrity and coherence of the instrument within the framework of teacher training programs used as a means of intervention, and supported the claim of at least a satisfactory level at validity and sensitivity.

Based on the results, it was deemed that the validity of the measuring was sufficiently high to be used in further analyses, e.g. for the assessment of programs' internal and external predictive validity in contextual variation, and also to be used as a means of intervention in the course of didactic observation and microteaching.

15.3 Phase III program predictive validation, a multivariate approach

15.3.1 Predicting success in student teaching from students' selection variables, rated and measured teaching behaviors and attitudes

Phase III of the study was designed for predictive validation in a longitudinal (1974-1988) ex-post facto empirical inquiry. The framework of the research used, and the results obtained, with the meta-level and substantive level strategies was presented earlier sections I and II. The main purpose of the final phase was the predictive validation of a Flanders-based (1965, 1970) study unit, a course of didactic observation and microteaching (cf. Heinilä 1977b), in the frame of the course's contextual setting, before and after the study degree program reform at the faculty (1978), by examining the variation of predictability in students' study success. A model was examined based on theories and assumptions of learning process which produces particular, "nondirective" teaching skills. The model was also based on program context,

presage, content, process and outcome relationships. It was hypothesized that students' study success, as well as their program evaluation results, were determined by their entry characteristics assessed in the selection procedure (intake stage 1. sum scores prior school success; stage 2. theory test, practice skills test, entry teaching behavior; and standardized intake sum scores). In addition, after two years measured and rated entry teaching behaviors and attitudes were assessed.

The framework of the research strategy used in this inquiry and its results obtained at the meta-level concerning the measuring instruments was presented earlier. The measured variables were created from faculty selection protocols, students' records and from the TV-recorded study material concerning rated and measured students process behaviors (PEIAC/LH-75 II, Scott's pi = .78). The empirical data did not meet the assumptions for the statistical treatment used, but the explanations and conclusion were, however, considered useful in curriculum evaluation.

When a replicated design and hierarchical regression were used to analyze the case studies, the theories of the learning process, subjects' characteristics, contextual setting, and program content effects on predictability were supported. Results from the multiple regression analyses showed that the predictor variables and study unit course intervention variables accounted for (MD) 34 % of the variance in students' study success in the final mark (theory and practice) on the course. The power of the selected regression model (R^2) variables to classify correctly students with low and high achievement scores ranged from 63 % to 77 % for the males and from 68 % to 78 % for the females. The effect of the other contextual factors such as the study program reform, changes in student selection procedures and also extra curricular effects - such prior experiences in teaching, changes in valuation of different teaching methods – would need to be considered to account for more of the variance (see e.g. Heinilä 1988, Silvennoinen et al. 1991, Telama 1970, 1990, Telama et al. 1988, Whitehead 1980).

The assumptions of subjects' characteristics and their contextual effects on predictability was supported, however. For example in a 1986/1988 case study ($n = 42$, $R^2 = .35$, $F(4.37) = 4.86$, $p = .003$), intake tests accounted for a main part (57%) of the explained variance in study success, final mark (theory and practice) - due to the combined effects of theory test scores (12%) (weight 30,7%), and of intake, stage 1. school success sum scores (8%) - as well as their interaction effects with students attitudes (7%) and entry teaching behavior (7%) (weight 11,5%). The role of gender, as determined at intake, was also supported, whereas the role of the intake practice skill (weight 55,5%) and of the total intake scores were not supported.

Predictability of study success was at higher level after the study reform and it was higher and more stable for the female students than for the males, judged on the basis of the classification power of the regression model variables. This was seen as a good indicator of the internal predictive validity of the program. The results also revealed gender differences, which were already identified in the results of students' intake scores ($r. 44$, $p < .01$). Thus gender effect was also revealed by analysing the correlations between predictor and

criterion variables. Therefore the regression analysis was conducted separately by gender. The results of this study paralleled findings of a large study project conducted in a natural setting (n = 117) by Varstala (1990). In this study, it was discovered that teacher's sex was important predictor of actual teaching behavior in physical education classes, and that the personal entry characteristics of students were the main predictors for the females. For the male students, attitudes about ideal P.E. teacher expectations and entry teaching behaviors were the most important predictors of success in the course.

Predictability for the main criterion – the assessment of students' ID index (acquisition of nondirective teaching skills) was at higher level than that of the other criteria. For example, in the case study intake course 1986/1988 the following models were extacted: for the male students (n = 21) R^2 = 66% F = (5,15) = 5.83, p < .01; power 86%, p< .01; for the female (n=21), R^2 = 30%, F (1,19) = 7.97, p< .01), power 81%, p< .05, and for the total sample (n = 42) R^2 = 30 %, F (3,38) = 5.43, p = .003) power 67 %, p< .03. Teaching behaviors or the "art of teaching", and its learning process were found to be very personal in nature, but could be based on a number of selected predictors such as the content of prediction models and their power. It was also found that the causes of study success for one student might not be the same for another students.

A overall conclusion based on the results obtained in Phases I, II and III of this study was that the program had quite good internal construct validity. The instrument for analysing interaction and its modification proved feasible both for research and for teacher training since it facilitated the employment of the intended behaviors, helped to teach observation, discrimination, and code patterns, and to create more indirect and flexible teaching behaviors. It also provided the main criterion for the program predictive validation.

The assumption concerning the external predictive validity of the program was also assessed in phase II and III results. However, the contextual effects proved to be confounding in explaining the findings. The Faculty selection procedures were of low predictive value and the total achievement level decreased significantly in two course populations after the study degree program reform in 1978. At this time, the course of didactic observation and microteaching was compressed to one two-week study unit, and the role of theory was weighted more than before; also the time reservation for the practice events in microteaching diminished (30 %). Predictability of study success, however, increased after the study reform. The program contextual variation also reflected strongly in students' (N = 283) program evaluation: students' feedback was positive, but less so than before, and overlapping aspects in faculty curriculum as well as time reservation problems were revealed. Moreover, the results of a discriminant analysis indicated that the measurements used for determining stability of students' ratings among the male, female and total two-tailed subpopulations appeared to be adequate, resulting in a discriminant validity of 95 %, 91 % and 77 % (p < .001) and placing group members correctly into their own course groups. This was also seen as an indicator of the validity of student program evaluation. The findings concerning student curriculum evaluation paralleled with results of faculty-wide study conducted by Rantakari and Tiainen (1983) and with results obtained in a 5-year follow-up evaluation

inquiry conducted by Telama, Rantakari and Rauhala (1988). Further research into interventions designed to develop students' social teaching skills within the preservice teacher training settings is, however, warranted. From the point of view of impacting of students teaching behaviors, the contextual factors would also need to be evaluated.

15.4 Strengths and weaknesses of the study

In spite of the many successful aspects of the study, it has several limitations. The most obvious is the limited scope of the empirical data. In the first phase of the project, the data consisted of boys' and girls' P.E. classes at three different grade levels taught by one male and one female teacher and covering four different areas of subject matter, a total of 24 lessons. This would have been a severe limitation if the purpose had been to make a generalizable description of what is happening in P.E. classroom interaction in Finnish schools. Such a description was not, however, the purpose of the study. For the purposes of initial testing of the developed instrument, the data was sufficient.

The major methodological problem of the study was the selection of the length of the time unit. Pilot studies had indicated that the three coded aspects (Clusters) had different natural rhythms. It was not possible to employ the much-used three-second-time unit due to the complexity of triple coding. *The decision to use a six-second arbitrary time unit to code all three clusters was a compromise made to allow the use of the same time unit in the simultaneous analysis of the whole process.* Naturally it was assumed that the aspects with slower tempo, such as the social form and the students' collective movement activity/passivity, would be reflected in various analyses as dominating features. This assumption was to be explored through a wide range of analyses.

Within these limitations, the study has contributed to the area of the study of instructional processes in P.E. classes. An observation instrument and a coding procedure were developed which went beyond the verbal orientation of most classroom interaction studies, and which incorporated features that reflected better the special characteristics of physical education classes as channel of communication and students' motor engagement time. The observation system PEIAC/LH-75 (Heinilä 1977a) and that the categories of the first and second clusters, as well as the six-second time unit have been used in a large study project connected with the assessment of teachers' and students' behavior and motor engagement time in school physical education classes (see e.g. Akkanen 1979, Reponen 1979, Varstala, Telama & Akkanen 1981, Varstala, Paukku & Telama 1983, Varstala 1990, Varstala 1996) and also in more than twenty postgraduate work (master and doctoral thesis) and projects conducted by the students who have taken this study unit course at the faculty and/or been research assistants one or more of the sub-studies reported in the above. In the application of the system to teacher training program, the use of the six-second time unit and double coding was found to be rational and giving the possibility to evaluate student's collective motor engagement time and process

behavior simultaneously. The criterion accepted for microlessons was 50 % motor engagement of effective class time. Obviously the use of this criterion in preservice teacher training from 1974 onward was reflected in further teaching behavior. Based on the results obtained in a large study (n = 406 lessons) it was noted that the Finnish students were physically active distinctly more often (about 50 % of effective classtime) than had been reported in studies conducted in the other countries, e.g. in Canada, France and U.S.A. (20 % - 30 %) (Piéron & Cheffers 1988; see also Varstala et al. 1981, Varstala 1996). Thus, on the basis of the work done, the instrument can be used to carry out more extensive and representative studies on the nature of interaction in P.E. classes. Since it was clearly demonstrated that content, the way of learning (cognitive, affective and psychomotor proprieties of the verbal and nonverbal communication) the subject matter area and the social form of classes were closely related to variation in the kind of classroom interaction, it would be useful to replicate the study with more subject matter areas and with more representative student samples.

Also the limitations of the ex-post facto program evaluation inquiry were highlighted in connection with program evaluation: the conclusions are probabilistic, reducing uncertainty but not totally eliminating it. However, for long-term investigations they are needed.

In the final phase, with the objective of program predictive validation, there was revealed a number of methodological issues related to ex-post facto inquiry that warrant further consideration. One of the main problems was connected to student selection procedure. E.g. when the test batteries are developed for student selection purposes and their parts are assigned certain weights, the objective is to obtain maximally high multiple correlations between predictors and criteria. Regression coefficients are not usually used as weights, however. There are several reasons for this. Some of them are purely technical, related to the linearity/nonlinearity of the relationship between predictors and criteria. Also, sampling errors are reflected in the validity coefficients, when the sum of weighted test scores are used in a new subject population. Other reasons are more substantive: certain characteristics may be considered so crucial for successful professional job performance that they cannot be compensated by other characteristics, e.g. teachers' clarity of presentation and creativity - "art of teaching" (see Flanders 1970, 270, 1987, 20, Rosenshine & Furst 1971,44). The results of this dissertation supported this insight. However, given the difficulty of identifying valid predictors of success, the selective admissions procedures might communicate program philosophy and improve the image of program, but obviously they cannot currently identify the "best", "ideal", prospective P.E. teachers (see Martens 1987).

Further, the study has highlighted the importance of the quality of teacher-student interaction and student-student interaction in physical education. Attention to this aspect is important if P.E. classes are to have the kind of impact on students' continued interest in physical activity. That is a major goal of P.E. teaching in our syllabuses (see e.g. Locke 1984, 5). Evaluation and feedback on the process of teacher education call for an appraisal of what changes have been made and what changes still need to be made in the

students' behavior so that the occupational demands of the changing teaching profession are fulfilled.

15.5 Implications for P.E. classroom teaching and teacher education

This study was carried out by a P.E. teacher who has also worked long in teacher education and who has a lifelong commitment to the improvement of teaching. The ultimate motivation for this study was thus to help develop P.E. teaching. Some recommendations can be made on the basis of the work done during the many years of the dissertation.

Teacher education programs in physical education cannot afford to focus too closely on one facet of personality, the psychomotor domain. The cognitive and affective aspects of physical education need to be fully appreciated by future and practicing P.E. teachers. The emphasis on the affective domain, which features prominently in PEIAC/LH-75, seems warranted on the basis of the extensive research on the Flanders system, but this should be ascertained specifically for physical education classes.

Through pre-service education, teachers should become familiar with the concept of indirect teacher behavior and its effects on classroom climate and interaction. This should be followed by a demonstration of how classroom interaction can be observed and analyzed. Didactic observation and microteaching in the pre-service training of future P.E. teachers has clearly indicated that this is possible and that it also opens a new perspective for students. Becoming critical about teaching amounts to moving from narrow "how to" questions to "what and why" questions, thus toward "reflective" teaching behavior.

15.6 Recommendations for further study

15.6.1 Observation instrument

During the more than fifteen-year period of the present study, a serious effort was made to explore a variety of issues and problems related to the empirical study of interaction in P.E. classrooms. However, several technical and methodological problems remain to be explored.

The results suggest that the following questions need to be addressed:

1. The development of rules for coding the verbal and nonverbal communication of teacher and students and of their sequences with a higher degree of specificity is desirable.
2. The optimum length of the coding interval needs careful consideration. A three-second interval might be appropriate in coding the first talk cluster, but a one-minute unit might be more reasonable in the other two clusters.

3. Rules for more decisive coding of students' collective movement activity and the forms of social access (categories II/3 and II/4) need to be developed.
4. The rules for coding students' collective passivity (II/7), waiting for turn, should be refined.
5. The rules for videotape recording need to be determined more exactly, specifying how the total situation is to be filmed.
6. Techniques for voice-recording need to be refined (e.g., using wireless throat microphones), with special attention to the problems of recording student talk.
7. The training of coders needs careful attention, with special emphasis on the content of training material so that sufficiently varied situations are presented to coder trainees.
8. Agreement controls carried out only at the beginning of coding are not enough to avoid systematic errors in coding. Recurring constancy control needs to be instituted.
9. The criteria for the selection of coders should take into account not only the cognitive but also the affective characteristics of rater candidates.

The empirical findings reported in this study concerning validity and sensitivity established clearly:

1. that in research work in connection with physical education several dimensions describing the influence patterns of the teacher are needed (see Cheffers 1973, 1977, Komulainen 1973),
2. that high frequencies of occurrence are not necessary prerequisites for the discriminant validity and sensitivity of the instrument. Nor should we be deterred from attempting to measure particular behaviors of interest from the point of view of theory on the ground that their occurrence is relatively infrequent,
3. that the aspect represented in the categories of the second and third clusters was found to be the dominating characteristic of the discrimination on the construct under study. It was related to the subject area of P.E., by analysing the formation of homogeneous groups based on factor-scores. But whether it must be so, is another question.

The feasibility of the instrument for different purposes needs to be considered more closely. It may be subscripted and postcripted so as to describe different patterns of students and teachers. The clusters can be used singly and/or inclusively, as was done in connection with teacher education programs, e.g. in microteaching (see Heinilä, 1977b, 1990). Double coding used to assessment of students' motor engagement time was important in P.E. class observation and feedback.

15.6.2 Curriculum evaluation

In summary, four significant points from the current investigation concerning curriculum evaluation warrant highlighting. First, one of the clear regular patterns found in this study was the interactive effect of the contextual setting.

Intake stage 1. (prior school success) had predictive value for study success in the study unit course, especially in the case of female students. After the study degree reform, when the courses of didactic observation and microteaching were completed, and the final note weighted more with theory (60 %), this issue was more visible.

Second, attitudes concerning students' expectations of "ideal" P.E. teacher characteristics were useful predictors of study success, provided that they expressed "student-centeredness" and support for freedom for expression, which was congruent with the original objectives of this course package based on ideas of Flanders (1965, 1970) and Rogers (1967). Such attitudes werestable over years (1974 – 1988), as shown by data covering four-intake course students. *However, the further studies of teaching behavior, students' attitudes, and intentionality relationships are warranted* (see e.g. Bain 1976, 1990, Flanders 1987, Martens 1987, Martin et al. 2001, Siedentop 1986).

Third, usually and in this data, as well the intake test practical skills were weak predictors of study success in this theory - weighted study unit. The results paralleled with results obtained in curriculum evaluation studies conducted in England (Whitehead & Hendry 1976, Whitehead 1980) and also with results obtained in Finland, Department of Physical Education (Silvennoinen, Laakso & Turunen 1991).

Fourth, in most of the several case studies conducted, students' entry teaching behavior, (teaching episode) had predictive value, for subsequent study success in the course, as expected. It was strongest in the intake samples - especially among female students - when its weight in intake procedure was high (20 % - 25 %), whereas in later cohorts, when its weight was diminished (15 % - 11,5 %) the predictability diminished. Its correlation with the total intake scores was not statistically by significant in most of the intake populations. Thus, it was largely a "chance" if a student with good entry teaching capacity was selected to the study program. It appeared that the selection procedure of the faculty would need subsequent evaluation of e.g. variation in weights used in test batteries.

It was also evident that didactic observation is a skill, related to the acquisition of reflective teaching behavior. However, its evaluation combined with subjects' microteaching course achievements had a confounding effect on the explanation of results. The learning of nondirective teaching skills demonstrated in microteaching setting would need further control and investigation in a follow-up study. From the point of view of influencing students' teaching behavior in a desired direction, the contextual factors such as curriculum program and its realization would also warrant further highlighting and evaluation.

Without the measurement of student entry characteristics or behavior any link between attitudes and teaching behavior is considerably weakened. To test the intention–teacher behavior link, research efforts examining this critical relationship would be important. Consequently, there exist other influences in the contextual setting, other than the investigated study unit, that impact on intentions and ultimately behavior that remain uncounted for, as well as

attitude formation itself (see e.g. Locke 1986, Telama 1970, 1990, Telama et al. 1988, Whitehead 1980).

Thus, the following points from the current investigation would warrant further highlighting:

- it would appear useful to have a variety of tests in student selection such as a teaching episode and test of students values and attitudes toward teaching,
- the predictive validity should be studied covering other study units and the overall performances of students in their final Physical Education Examinations,
- a research design with different students characteristics, i.e. student groups selected with different criteria, might bring further light to the problem of the validity of the curriculum program covering different study units and selection procedures
- further research into interventions designed to develop teaching behavior of student teachers within the pre- and in-service settings is warranted.
- finally, investigations of other important curriculum objectives as e.g. student's intentions to teach other social skills and creativity would certainly complement and expand current findings on the predictability of the development of reflective teaching behavior.

In summary, the observation instrument and the coding procedures still would need refinement. This is to be expected. Flanders (1987, 242) pointed out that "the fact that teaching is a complex social process, hard to define and evaluate, does not mean that all evidence is useless simply because it is incomplete. The tools and techniques to establish criteria of teaching effectiveness are crude, but they can be improved only by further experimentation and development". One of the weaknesses in current research procedures is, as Cheffers (1990) has pointed out, that repeated research with constant problems are rare. In this long-term research project this kind of design was used. This suggested approach produced useful knowledge for the implementation of teaching-learning conditions - for preparing P.E. teachers who are capable of technical, practical and critical reflection.

Laakso observed in the early 1980s (1984, 131-134) that "the in-depth study of class interaction in P.E. classes in Finland has only begun". When this long-term study was mounted, it was done in order to get a better grasp of the conceptual and methodological issues and problems in such type of investigation. It is to be hoped that this "beginning" has been greatly extended and intensified and the study will grow and flourish. The course program information and results of the evaluation study has been presented in seven different languages and the results have been presented in a number of contexts - Finnish, Swedish, French, German, English, Portuguese and Spanish. The developed course package was the study curriculum of the faculty of the University of Jyväskylä Finland from the year 1974 to 1991, and from the year 1995 the course of didactic observation and microteaching has been conducted in another form and the new program has also been evaluated (Heikinaro-Johansson & Varstala, 2000). Thus there is a continuing interest in studying

what contributions interaction analysis might be able to make to teacher education and practice teaching. The author shares Barrett's (1996, 144) view, concerning curriculum preparation, that "we need strategies, which combat curricular fragmentation and faculty dissociation and which promote convergent strategies of planned integration connected to the real work of teachers in a long overture". The complexity of the interaction in P.E. classes is great but, with continuous study, we can be confident that our knowledge base will keep increasing at a steady face.

16 YHTEENVETO

Vuorovaikutusprosessin analyysimenetelmän kehittäminen liikuntatilanteisiin, sen soveltaminen liikunnanopettajan koulutusohjelmaan sekä ohjelman arviointi

Vuorovaikutusta koskevat yleiskäsitykset ovat lähtöisin useista eri lähteistä. Tämän työn teorialähtökohdat ovat sosiaalipsykologiasta, joka 1970 luvulta alkaen on ollut vallitseva suuntaus myös liikunnanopetuksen ja opettajankoulutuksen tutkimuksessa. Kaiken keskipisteenä on pienryhmätoiminnassa, kuten opetuksessa, vuorovaikutusyksikköjen luoma ja säätelemä sosiaalinen järjestelmä, joka sisältää ainakin seuraavat osatekijät: 1) joukon yksikköjä, jotka ovat vuorovaikutuksessa keskenään, 2) joukon sääntöjä, jotka säätelevät sekä yksiköiden orientoitumista että vuorovaikutusta itseään, 3) ajallisesti muotoutuneen vuorovaikutussysteemin tai prosessin sekä 4) ympäristön, jossa järjestelmä toimii ja jonka kanssa tapahtuu systemaattista vuorovaikutusta (Heinilä 1974, Parsons 1968).

Opetusprosessin tieteellisyyteen perustuva hallinta oli asetettu tavoitteeksi uudistuvassa liikunnanopettajakoulutuksessa (Komiteamietintö 1975: 75). Tutkimus on voinut osoittaa interaktiomenetelmien opettajakoulutuksen välineinä tarjoavan entistä paremmat mahdollisuudet tämän tavoitteen saavuttamiseksi, ja samalla teorian ja käytännön integroimiseksi.

Korkeakoulupedagogiikka, samoin kuin myös uusimuotoisten opettajankoulutusohjelmien kehittäminen, edellytti niiden sisällöllisten ja laadullisten peruselementtien kontrollointia ja evaluointia. Objektiivisen tiedon hankintaan tarjosivat juuri interaktioanalyysimenetelmät ja uudistunut teknologia entistä paremmat mahdollisuudet. Samalla ne mahdollistivat tavoitekäyttäytymisen informoinnin, operationalisoinnin sekä opetuskäyttäytymisen analysoinnin ja arvioinnin, jotka juuri ovat mm. opetusharjoittelun keskeisiä elementtejä. Näiden elementtien tuntemus oli 1970-luvulla liikunnanopettajakoulutuksessa perin vähäistä.

Tämä väitöskirjatyö koostuu useista 30 vuoden aikana tehdyistä tutkimuksista ja työn alussa esitetyistä julkaisuista (Preface) (Heinilä 1970 – 1997).

Tutkimusprojekti ja siihen perustuva opettajankoulutusohjelma käynnistyi 1970-luvun alussa Jyväskylän yliopiston liikuntakasvatuksen laitoksella. Sen tarkoituksena on ollut edistää liikuntakasvatusta luomalla entistä parempia mahdollisuuksia liikunnan opetusprosessin mittaamiselle, analysoinnille ja arvioinnille sekä uusimuotoisten opetusharjoitteluohjelmien kehittämiselle ja evaluoinnille siten, että ne palvelisivat entistä tehokkaammin korkeakoulun liikunnan opetukselle asettamien ammatillisten tavoitteiden saavuttamista.

Osa I

Tutkimus on kolmivaiheinen. Ensimmäisessä vaiheessa laadittiin esitettyyn laajaan kirjallisuuskatsaukseen perustuen ongelman asettelun ja tilastoanalyysien perustaksi malli, joka kuvasi meta- ja substantiivisella tasolla tapahtuvaa tutkimusta sekä taustan, panoksen, prosessin ja tuotosten välisiä yhteyksiä. Aluksi oli tavoitteena kehittää systemaattiseen observointiin perustuva liikunnan vuorovaikutusprosessin analyysijärjestelmä. Keskeisenä ongelmana oli mittavälineen analyysimenetelmien kehittäminen, sekä sen validiteetin ja reliabiliteetin, soveltuvuuden ja erottelukyvyn arviointi. Kehitetty observointiväline perustuu Flandersin teoriaan (1965, 1970) sekä liikunnan opetustapahtuman empiiriseen tutkimukseen (Heinilä 1971, 1974). Opettajan vaikutustapaa selittäviksi ulottuvuuksiksi asetettiin seuraavat käsitteelliset perusulottuvuudet: 1) tavoitteellisuus (oppilaan kannalta), 2) opettajan auktoriteetin aste (opettajan aloitteisuus/oppilaan aloitteisuus), 3) oppilaiden sosiaalisen toiminnan vapaus/sosiaaliset kontaktit (kollektiivinen aktiivisuus – passiivisuus) sekä 4) sosiaalimuoto (työn ja vastuun jako). *Nämä käsitteet on tarkoitettu välineiksi ajateltaessa oppilaan itsenäisyyden ja itseohjautuvuuden kehittämistä.* Luokitusjärjestelmän klusterit ovat: 1) opettajan ja oppilaiden puhe (12 kategoriaa), 2) oppilaiden kollektiivinen liiketoiminta (8 kategoriaa) ja 3) sosiaalimuoto (7 kategoriaa). Kolmiulotteisessa luokituksessa on otantayksikkönä 6 sekunnin aikaväli.

Menetelmällisen tutkimusvaiheen aineisto käsitti kuuden koulutetun observoijan luonnollisissa ja tv-nauhoitetusta aineistosta kolmasti observoimat 24 eri sukupuolta olevien opettajien (ja oppilaiden) kolmen luokkatasoryhmän ja neljän liikunnan oppiainesalueen tunteja. Aineistoa analysoitiin kuvailun ja vertailun tasolla: profiilit, matriisit, indeksit, r-korrelaatiomatriisit, varimax-rotatoidut faktorit, homogeeniset rakenneryhmät, niiden eroavuus ja eroja aiheuttavat tekijät (Heinilä 1976, 1983). Mittauksen objektiivisuutta tutkittiin eri klusterien osalta analysoimalla mm. Scottin pii-kertoimien varianssia. (ANOVA ja t-testit). Kendallin W-kerrointa käytettiin eri kategorioitten luokituksen yhdenmukaisuuden arvioinnissa. Eri luokittajien observoinnin eroavuutta ja eroa aiheuttavia tekijöitä tarkasteltiin käyttäen erotteluanalyysia (Heinilä 1980).

Saatujen tulosten perusteella voitiin todeta, että kehitetty, Flandersin vuorovaikutusprosessin analyysijärjestelmään ja tutkimuksiin (Heinilä 1971, 1974, 1976, 1977a) perustuva liikunnan vuorovaikutusprosessin analyysijärjestelmä PEIAC/LH-75 soveltuu käytettäväksi erilaisten liikuntatuntien observointiin ja analysointiin, ja että sen rakennevaliditeetti täyttää tietyn tason luotettavuusvaatimukset ja että se omaa riittävän herkkyyden suoran ja epä-

suoran opetuksen erotteluun. Saatua tulosta tukivat Reposen (1979) esitutkimuksen (n = 44) antamat tulokset. Lisäksi tutkimustulosten perusteella voitiin luoda liikunnan opetustapahtuman toimintoja kuvaava malli. Sen pääkomponentteina, jotka rajaavat opettajan ja oppilaiden toimintaa, ovat opetustapahtuman tavoitteellisuus ja sosiaalinen rakenne. Tätä mallia on käytetty seuraavassa vaiheessa viitekehyksenä sovellettaessa kehitettyä järjestelmää opettajankoulutusohjelmaan interventiostrategian ja opetusprosessin peruselementtien kuvaamiseksi.

Osa II

Tutkimuksen toisessa vaiheessa, sovellettaessa PEIAC/LH-75 järjestelmää liikunnan opettajankoulutukseen, oli tarkoituksena saada aluksi tietoa pienoisopetuskurssin sisältöä, muotoa ja ajoitusta koskeviin kysymyksiin. Tämän opetusohjelman evaluointitutkimuksen ongelmana oli arvioida ja verrata kahta pienoisopetuksen opetuspakettia, joissa tavoitteena oli kehittää opettajakokelaiden sanallista epäsuoraa opetuskäyttäytymistä. Tavoitteiden prosessin ja tuotosten välistä kongruenssia tutkittiin. Opetuspaketit erosivat toisistaan seuraavien elementtien suhteen: 1) informaatio tavoitekäyttäytymisestä, opetusmallit (kirjallinen vs. AV-materiaali), 2) teoriaopetuksen ajoitus (ennen kurssia vs. kurssin aikana), 3) ryhmäkoko (5 – 10), 4) mikro-opetustuntien kesto (5 – 10 min), ja 5) mikro-opetustuntien lukumäärä (2 – 3).

Tutkimuksen kohdeaineiston muodostivat vuonna 1974 ja 1976 järjestetyt kurssit (15 t luentoja, 30 t harjoittelua) ja koehenkilöinä olivat vuosina 1971 (n = 48) ja 1974 (n = 74) opintonsa tiedekunnassa aloittaneet mies- ja naisopiskelijat. Observointivälinettä (PEIAC/LH-75 II) käytettiin muunnettuna sisältäen kaksi klusteria: 1) opettajan ja oppilaiden puhe sekä 2) oppilaiden kollektiivinen liiketoiminta sisältäen 16 ja 2 kategoriaa. Kaksoiskoodaus tapahtui 6 sekunnin aikavälein. Mittaväline mahdollisti tavoitekäyttäytymisen (opetusmallien) operationalisoinnin ja tulosten analysoinnin ja vertailun.

Tutkimuksen aineisto (54 ja 148 pienoisopetustuntia) kerättiin TV-nauhoitteista. Reliabiliteetti arvioituna käyttäen Scott'n pii-kerrointa oli I .79, II .98. Kategoriafrekvenssejä, indeksejä ja opiskelijoiden kurssin evaluointituloksia, jotka oli kerätty kyselylomakkeella (58 osiota, 1 – 5 asteikko), verrattiin käyttäen varianssianalyysiä ja t-testiä (ANOVA), sekä Mann-Whitney U-testiä että Khin neliötestiä. Tulokset tukivat asetettua olettamusta jälkimmäisen kurssimuodon tehokkuudesta ja soveltuvuudesta. Sekä prosessievaluoinnin vertailutulokset että opiskelijoiden palaute antoivat tarvittavaa tietoa kurssin lopullisen interventiostrategiamallin rakentamiseksi sekä seuraavan evaluointitutkimuksen vaiheiden suunnittelemiseksi.

Seuraavassa tutkimusvaiheessa keskityttiin kehitetyn didaktisen observointi- ja pienoisopetuksen kurssin, sen interventiostrategian ja peruselementtien luotettavuuden ja sopivuuden arviointiin. Aluksi tutkittiin opettajankoulutukseen sovelletun PEIAC/LH-75 II observointivälineen validiteettia, sekä sen soveltuvuutta tavoitekäyttäytymisen "epäsuorien" opetusmallien toteutuksen arviointiin. Aineiston muodostivat nauhoitetut pienoisopetustunnit 1, 2 ja 3 (n = 74, 222 t ja n = 42, 126 t). PEIAC/LH-75 II järjestelmä ja observointiväline

osoittautuivat reliaabeliksi, validiksi ja soveltuvaksi käytettäväksi kurssilla epäsuorien opetusmallien toteuttamisen arviointiin sekä opetusprosessin tutkimuksiin. Akkasen (1979) luonnollisissa tilanteissa (n = 8) suorittama esitutkimus, jossa käytettiin PEIAC/LH-75 II järjestelmää ja kehitettyjä suoran ja epäsuoran opetuksen malleja, tuki osaltaan saatua tulosta. (Ks. Heinilä 1977b, 1990, 1992).

Toiseksi tutkittiin opiskelijoiden opetuksen arvioinnissa valintakokeen opetustuokiokokeessa (3,5 min) ja pienoisopetuksen kurssin opetustuokiokokeessa (5 min, kontrolli) käytetyn mittavälineen (4 osiota, 1-6 asteikko) reliabiliteettia ja validiteettia. Arvioinnin kohteena ovat (1) esitystapa, (2) sisällön ymmärrettävyys, (3) kommunikointi ja vuorovaikutus sekä (4) luovuus. Mittaria käytettäessä kiinnitetään huomio siihen kuinka selvästi "oppilaille" pystytään esittämään annetun tehtävän tavoitteet, sisältö ja toimintaohjeet sekä kuinka "oppilaiden ja opettajan välinen vuorovaikutus tilanteessa tapahtuu, ja kuinka osuvasti tehtävän kehittelyvaiheessa käytetään omaa harkintaa ja mielikuvitusta", kuten koehenkilöille annetussa kokeen esitteessä sanotaan. Kaksi ulkopuolista koulutettua tarkkailijaa arvioivat koehenkilöiden ("opettajien") suoritukset. Aineistona käytettiin opiskelijoiden valintakokeessa saamia summapistemääriä sekä kaksi vuotta myöhemmin ennen pienoisopetuksen kurssia (kontrolli) kerättyä TV-nauhoitettua aineistoa (n = 42, n=75). Nauhoitukset käsiteltiin sekä arviointia että observointitekniikkaa käyttäen. Mittavälineen erottelukykyä ja luotettavuutta arvioitiin käyttäen lisäksi aineistona neljän eri kurssiryhmän opiskelijoiden (n=205) opetuksen arvioinnissa saatuja tuloksia. Mittausten luotettavuutta todistaa osaltaan tulosten pysyvyys: kahden vuoden välein mitatut opetustuokiokokeiden summapistemäärien väliset Pearson-korrelaatiokertoimet olivat useimpien kohdeaineistojen osalta tilastollisesti merkitsevät. Todettakoon, että arvioijat olivat näissä tilanteissa eri henkilöitä. *Tulokset osoittivat mittavälineen omaavan riittävän luotettavuuden tason ja soveltuvan käytettäväksi valintakokeen ja pienoisopetuksen kurssin opetustuokiokokeen (kontrolli) arviointivälineenä ja niihin liittyvissä tutkimuksissa.*

Myös opiskelijoiden asenteiden, "ihanneliikunnanopettajan" luonteen ominaisuuksien arviointiin käytetyn, Flandersin (1965, 1970) ja Rogersin (1967) esittämiin ideoihin perustuvan kaksiulotteisen kyselylomakkeen (20 osiota 1-6 asteikko) reliabiliteetti ja validiteetti arvioitiin. Aineistona monimuuttuja-analyysissä olivat neljän kohderyhmän mies- ja naisopiskelijoiden (n = 204) arvioinnin faktoriulottuvuudet. Neljän faktorin ratkaisun korrelaatioiden perusteella laskettiin Cronbachin alfan vaihteluväli (.56 - .92). "Ihanneopettajaa koskevien odotusten" voitiin todeta olleen pääulottuvuudella (52 %) oppilaskeskeinen – opettajakeskeinen" vakaat riippumatta sukupuoli- tai kohderyhmä- ja aikatekijöistä. Sen sijaan mm. opettajan osallistumista kuvaavalla ulottuvuudella näkyi huomattava muutos 1980-luvulla ja erityisesti naisilla. Ilmeisesti se oli "aerobicin" vaikutusta. *Tulosten perusteella voitiin todeta mittavälineen luotettavuus sekä sen soveltuvuus käytettäväksi kurssin interventio- sekä tutkimusvälineenä (ks. myös Heinilä 1988, 1992b).*

Tutkimusprojektin viimeisessä vaiheessa päätavoitteena oli kehitetyn opintoyksikön (2 opintoviikkoa) ohjelman sisäisen ja ulkoisen ennustevaliditeetin arviointi. Ongelmanasettelun ja tilastoanalyysien perustaksi laadittiin

tutkimuksiin perustuen malli, joka kuvasi ohjelman taustan, opiskelijoiden taustatekijöiden, kurssiohjelman, prosessin ja tuotosten välisiä yhteyksiä (Dunkin 1987). Tutkimukselle asetettiin seuraavat ongelmat: 1) selvittää millaisia yhteyksiä ilmenee didaktisen observoinnin ja pienoisopetuksen kurssin opiskelijoiden opintomenestyksen ja taustamuuttujien, kuten valintakoetestien sekä kaksi vuotta myöhemmin mitattujen opetuskäyttäytymisen testitulosten ja asenteiden välillä, 2) selvittää näiden yhteyksien ennustettavuuden tason ja koulutusohjelman taustatekijöiden (eri kohderyhmien ja tiedekunnan tutkinnon uudistuksen) välisiä yhteyksiä sekä 3) tutkia, missä määrin laadittu hypoteettinen kurssin opintomenestyksen mallin ennuste on uskottava ja tilastollisesti todennettavissa.

Tutkimuksessa neljältä kohderyhmältä kerätty tiedosto (n = 205) edusti noin 40 % kaikista 1976 –1988 kurssiaineistoista (1974/1976, 1976/1979, 1977/1980, 1986/1988). Valintakokeessa hyväksyttyjen ja kurssin suorittaneiden aineisto-osuus oli 85 – 90 %. Opintomenestyksen ennustettavuuden arvioinnissa käytettiin kriteereinä viittä tavoitesidonnaista muuttujaa: kurssilla mitattu (PEIAC/LH-75 II) 1) F-1:den ja 2) ID-indeksin toisen ja kolmannen pienoisopetustunnin summapistemäärä sekä 3) teorian, 4) käytännön ja koko 5) opintokurssin arvosana (1 – 3). Ennustemuuttujina käytettiin opiskelijoiden taustatekijöitä kuvaavia tietoja, jotka oli kerätty tiedekunnan valintakokeen yhteydessä: 1) ensimmäisen vaiheen summapistemäärä (mm. aikaisempi koulumenestys), 2) toisen vaiheen teoriatesti, (3) käytännön taitotesti, (4) opetustuokiokokeen summapistemäärä, (5) valintakokeen standardisoitu kokonaispistemäärä sekä kaksi vuotta myöhemmin (6) arvioitu opetustaito (kontrolli, 5 min), osiot 1-4 ja (7) mitattu (PEIAC/LH-75 II) opetustuokion (kontrolli) pistemäärä F1 ja ID-indeksi, sekä lisäksi (8) opiskelijoiden asenteet, heidän "ihanneliikunnanopettajansa" luonteen ominaisuuksien arviointitulokset (faktoripistemäärät F1 – F4). Arvosanat, kuten myös valintakoetulokset, oli otettu tiedekunnan pöytäkirjoista ja opintorekisteristä.

Tutkimuksen aikaisemmassa vaiheessa oli esitetty arvio käytettyjen mittavälineiden luotettavuudesta ja niiden erottelukyvystä, jota oli testattu kohderyhmien sekä varianssien alaryhmien homogeenisuutta arvioitaessa. Aluksi koko aineistolle laskettiin keskiarvot, hajonnat, vaihteluvälit ja validit havaintomäärät. Varsinainen tilastollinen käsittely aloitettiin tarkastelemalla yksittäisten tausta- ja opintomenestysmuuttujien välisiä yhteyksiä. Tämä tapahtui Pearson-korrelaatiomallilla. Teoreettiseen tutkimukseen perustuvan ennuste- ja selitysmallin testauksessa käytettiin valikoivaa regressioanalyysiä, jolloin muuttujien keskinäiset riippuvuudet otetaan huomioon, sekä erotteluanalyysia ennustettavuuden voimakkuutta arvioitaessa. Opiskelijoiden opintomenestystä kuvaavat muuttujat oli dikotomisoitu. Nämä analyysit tehtiin vuosikurssikohtaisesti ja erikseen opiskelijoiden sukupuolen mukaan. Saatuja kohderyhmien tuloksia verrattiin eri kriteerien ja eri sukupuolten osalta.

Käytettäessä toistettua tutkimusasetelmaa ja valikoivaa regressioanalyysiä muodostui opintomenestyksen ennustettavuudesta seuraavanlainen kuvaus: *teorian osalta* tuli yhteiskorrelaatio tilastollisesti merkitseväksi miehillä vain yhdessä aineistossa neljästä, kun taas naisilla kolmessa neljästä. Naisopiskelijoiden osalta selitysprosentit vaihtelivat 33 %:sta - 50 %:iin ja saatujen ennustemallien

voimakkuuden vaihteluväli oli 71 % - 84 %. Regressiomallit koostuivat näissä tapauksissa 1 – 3 muuttujasta, jolloin sekä valintakokeen muuttujat, 1. vaiheen summapistemäärä (mm. koulumenestys), 2.vaiheen teorian testi – (jälkeen tutkinnonuudistuksen) sekä arvioitu opetuskäyttäytyminen (opetuksen selkeys) ja asenteet olivat valikoituneet ennustemalliin. Miehille valikoituneen ainoan ennustemallin selitysprosentti oli 18 % ja sen ennustemuuttuja oli valintakokeen teoriatesti.

Käytettäessä *käytännön arvosanaa kriteerinä* oli miesten ja naisten opintomenestyksen ennustettavuus tasaisempaa: miehille valikoitui neljä ennustemallia ja naisille kolme neljästä, joiden selitysprosentit olivat keskimäärin miehillä MD 38 % ja naisilla MD 37 %. Myös mallien erotteluvoimakkuus oli jokseenkin samalla tasolla: miehillä MD 68 % ja naisilla MD 65 %.

Loppuarvosanan, opintomenestyksen, varianssia selitettäessä olivat selitysprosentit jonkin verran edellistä matalammalla tasolla. Yhteiskorrelaatio tuli miehillä merkitseväksi kaikissa neljässä kohdeaineistossa ja naisilla kolmessa, ja mallien selitysprosentit olivat keskimäärin miehillä MD 33 % ja naisilla MD 34 %. Ennustemallien muuttujien erotteluvoimakkuus oli naisilla hieman korkeampi (MD 76 %) kuin miehillä (MD 71 %).

Naisten ja miesten ennustemallit poikkesivat toisistaan muuttujien, niiden lukumäärän sekä selitysprosenttien suhteen. Naisilla yleensä, esimerkiksi vuoden 1988 (n = 21) aineistossa, loppuarvosanan malli koostui kahdesta muuttujasta, jolloin valintakokeen 1. vaiheen summapistemäärä (aikaisempi koulumenestys) ja arvioitu opetuskäyttäytymisen (luovuus ja esityksen selkeys) valikoituvat ennustemalliin, kun taas miehillä ennen kaikkea asenteet ("ihanneliikunnanopettajan" ominaisuudet) sekä arvioitu ja mitattu opetuskäyttäytyminen (luovuus) olivat tärkeimpinä loppuarvosanan ennustajina. Mainitun kohderyhmän kaikkien opiskelijoiden (N = 42) osalta saatiin ennustemalli (R^2 = .35, F(4,37) = 4.86, p = .003), jossa valintakoemuuttujat selittävät huomattavan osan (57 %), toisen vaiheen teoriatestin (12 %) ja ensimmäisen vaiheen summapistemäärän (8 %) sekä näiden yhteisvaikutuksen ja edelleen valikoituneiden opiskelijoiden asenteiden (7 %) ja arvioidun opetuskäyttäytymisen (7 %) yhteisvaikutuksen ansiosta.

Saadut tulokset tukivat asetettua olettamusta opintomenestyksen ennustettavuudesta ja ennusteen tilastollisesta uskottavuudesta. Malleihin valikoituneet opiskelijoiden kognitiivinen kapasiteetti, asenteet ja aikaisempi opetuskäyttäytyminen yhdessä tukivat tehtyä olettamusta epäsuoran opetusmallien oppimisen ennustettavuudesta kurssilla joka perustui teoriaan (Flanders 1965, 1970, Heinilä 1977b, 1992, Rogers 1967) sekä kurssin sisäisen ennustevaliditeetin arviointia.

Tulokset osoittivat myös, että käytettäessä kehitettyä vuorovaikutusprosessin analyysijärjestelmää (PEIAC/LH-75 II), opiskelijoiden mitatun prosessikäyttäytymisen arvioinnissa kriteerimuuttujien (F-1 ja ID-indeksi "epäsuora opetuskäyttäytyminen") ennustettavuus oli korkeammalla tasolla kuin muiden tässä tutkimuksessa käytettyjen kriteerimuuttujien, loppuarvosanojen, jotka oli yhdistetty teorian ja käytännön opintosuorituksista. Mm. esimerkkiaineistossa (1988 n = 42) ID-indeksin selitysprosentti oli miehillä 66 % ja naisilla 30 %. Näiden mallien, jotka koostuivat useista (1 – 5) muuttujista,

erotteluvoimakkuus oli sekä miehillä että naisilla varsin korkea 86 % ja 81 %. Voitiin myös todeta, että alhaisen pistemäärän saaneiden miesten (n = 21) osalta erottelufunktion voimakkuus oli jopa 100 %, kun taas opinnoissa menestyneiden osalta vain 70 %, kun taas naisilla (n = 21) ennustettavuus sekä menestyneiden että vähemmän menestyneiden osalta oli tasaista (81 % - 82 %). Tämä johtui ilmeisesti ennustemuuttujien erilaisuudesta. Myös loppuarvosanan ennustettavuudessa näkyi samankaltainen ero sukupuolten välillä. Näihin mitatun opetustaidon ennustemalleihin valikoituivat useimmiten aikaisempaa opetustaitoa ja varsinkin miehillä myös asenteita edustavat muuttujat.

Tutkimuksessa voitiin todeta, että valintakokeen *opetustuokion* sekä kaksi vuotta myöhemmin pidetyn opetustuokiokokeen tuloksilla oli ennustearvoa epäsuoran opetuksen ymmärtämisen sekä opetusmallien toteuttamisen – ts. tavoiteoppimisen kannalta. Todettakoon vielä, että mittaustulosten pysyvyys, niiden välinen Pearson-korrelaatio (r = .52) oli tilastollisesti merkitsevä mm. vuoden 1988 kohderyhmän aineistoissa. Yhtyneenä opiskelijan kognitiiviseen kapasiteettiin ja asenteisiin opetuskäyttäytyminen tuki osaltaan myös koko opintomenestyksen ennustettavuutta. Tämä näkyi selvimmin, kun kriteerinä käytettiin mitattua opetuskäyttäytymistä (ID-indeksi). Aikaisempi opetustaito, kuten valintakokeen opetustuokion summapistemäärä ja kaksi vuotta myöhemmin arvioitu opetustuokion (kontrollin) osio 1 "selkeys" ja osio 4 "luovuus" sekä myös mitattu opetuskäyttäytyminen (ID-indeksi) yhtyneinä positiivisiin asenteisiin (F-1), olivat hyviä opintomenestyksen ennustajia tällä kurssilla. Myös muissa tutkimuksissa, mm. kokeellisissa tehokkuustutkimuksissa, joissa kriteerimuuttujana on ollut observointiin ja arviointiin perustuva mitattu opetuskäyttäytyminen, esityksen selkeyden pistemäärien korrelaatioiden on, todettu olevan tilastollisesti merkitsevät (Rosenshine & Furst 1971, 44).

Tämän tutkimuksen yhteydessä voitiin lisäksi todeta, että opetustuokiokokeessa arvioitu "opettajan" esityksen selkeys, erityisesti miesten osalta, oli parantunut aikaa myöten ja ero oli tilastollisesti merkitsevä 1970- ja 1980-luvun aineistojen (n = 205) vertailussa. Luonnollisesti ilmaisua, samoin kuin muitakin opetustaitoja voi harjoituksen avulla parantaa. Näin on ilmeisesti tapahtunut kohdeaineiston analyysitulosten perusteella, kun tämä valintakoekriteeri on tiedostettu. On hyvä, että tätä opetuksen tehokkuuteen liittyvää sekä epäsuoran opetuskäyttäytymisen oppimisen kannalta tärkeää ominaisuutta arvostetaan ainakin opiskelijoiden taholta. Tässä tutkimuksessa voitiin kuitenkin todeta, että sen korrelaatio valintakokeen kokonaispistemäärään ei ollut tilastollisesti merkitsevä – varsinkin kun sen painokerroin oli 11,5 %. On siis sattuma, jos tutkimusaineistossa on tämän kokeen perusteella hyvän opetustaidon omaavia koehenkilöitä. Ilmeisesti myös opetustuokiokokeen painoarvon lisäämistä valintakokeessa tulisi harkita.

Opiskelijoiden asenteiden merkitys "ihanneliikunnanopettajan" ominaisuuksien arvioinnissa niin opetuskäyttäytymisen kuin myös koko opintomenestyksen ennustettavuuden kannalta näkyi selkeästi. Lisäksi havaittiin, että faktorianalyysin perusteella saatu pääulottuvuus (F-1) "oppilaskeskeisyys"/ "opettajakeskeisyys" (ts. aitous, joustavuus, sopivuus ja asianmukaisuus) oli varsin vakaa eri kohderyhmissä yli kahden vuosikymmenen ajan molempien sukupuolten kohdalla. Tämä osoitti samalla kurssin opiskelijoiden asenteiden

yhdenmukaisuutta ja tavoitteisuuden selkeyttä. Saadut tulokset tukivat osaltaan opintomenestyksen ennustettavuutta. Tavoitteisuus on juuri Flandersin (1965, 1970) teoriaan perustuvassa PEIAC/LH-75 (Heinilä 1974, 1976) järjestelmässä keskeinen opetuksen tehokkuuden hypoteellisen mallin osatekijä. Toisaalta myös muissa tutkimuksissa on voitu osoittaa, että asenteet ovat merkittävin opettajan tavoitteisuuden ennustaja, mikä todettiin mm. tutkimuksessa, jossa kysymys oli liikuntatuntien motorisen aktiivisuuden määrän arvostamisen ennustettavuudesta (Martin, Kulinna, Eklund & Reed 2001). Ilmeisesti liikuntatieteelliseen tiedekuntaan hakijoiden kasvattaja-asenteita ja arvoja mittaavien testien käyttöönottoa olisi syytä harkita - mihin suositukseen päätyivät myös Silvennoinen, Laakso ja Turunen (1991) valintakokeen ennustevaliditeetin arvioitiin kohdistuneen tutkimuksensa perusteella.

Tutkimuksessa todettiin lisäksi, että *didaktisen observointikurssin ennustearvo pienoisopetuksen kurssin opintomenestyksen ennustettavuuteen oli korkea.* Mutta kun kurssit tutkinnon uudistuksen yhteydessä yhdistettiin, molempien opintoyksiköiden oppisaavutukset arvioitiin vain yhdellä loppuarvosanalla (50 % ja 50 %) ja teorian osuutta painotettiin molemmissa arvosanoissa (60 %). Tämä luonnollisesti heijastui ennustettavuuteen. Tällöin myös vain "hyvän" opetustaidon ("art of teaching") omaavat saivat ilmeisesti liian alhaiset arvosanat.

Tuloksien perusteella voitiin myös todeta, etteivät valintakokeen käytännön taitojen testi, eikä sen standardisoitu kokonaispistemääräkään tukeneet opintomenestyksen ennustettavuutta eivätkä juuri valikoituneet ennustemalleihin. Käytännön kokeella oli eräissä malleissa valikoituessaan negatiivinen korrelaatio kriteerimuuttujaan. Lisäksi mitatun opetustaidon opintomenestyksen selittäminen tapahtui paremmin taustamuuttujien kuin valintakokeiden antaman informaation avulla. Kun lisäksi todetaan, että naisopiskelijoiden osalta aikaisempaa koulumenestystä koskevat tiedot selittävät kurssin loppuarvosanan varianssista enemmän kuin muut ennustemuuttujat, jää pääsytutkinnon kokonaispistemäärälle tiettyjä painokertoimia käytettäessä epämääräinen, tilastollisesti ei merkitsevä kurssin opintomenestyksen ennustearvo. Myös muissa, koko tiedekunnan eri opintoyksiköitä kattavissa tutkimuksissa on päädytty samankaltaisiin johtopäätöksiin koskien mm. valintakokeen ennustevaliditeettia ja mm. opetustaidon loppuarvosanan ennustettavuutta (ks. mm. Silvennoinen, Laakso & Turunen 1991). Saavutettujen tulosten voidaan todeta olevan samansuuntaiset kuin Englannissa liikunnanopiskelijoiden valintakokeiden validiteettitutkimuksessa (Whitehead & Hendry 1976, 129 – 130, Whitehead 1980). Siinä valintakokeiden käytännön testipisteiden ei todettu olevan yhteydessä loppuarvosanaan eikä myöhempään ammattimenestykseen. Saavutettujen tulosten perusteella on aihetta harkita muunlaisten valintamenetelmien käyttöä.

Tämän tutkimuksen tulokset tukivat olettamusta koulutusohjelman taustatekijöiden yhteydestä tietyn opintokurssin opintomenestyksen ennustettavuuteen ja siinä esiintyviin vaihteluihin. Taustatekijöiden, kuten tiedekunnan valintakokeen testipatteristojen painotusmuutosten sekä tutkinnonuudistuksen yhteydessä tehtyjen, koko koulutusohjelmaa ja eri opintoyksiköitä ja niiden arviointia koskevien muutosten, todettiin heijastuvan myös didaktisen observointi ja pienoisopetuskurssin opintomenestyksen ennustettavuuteen. Mm.

kurssiryhmissä, joiden valintakokeissa teoriaa oli painotettu paljon (1974, 1986), ennustettavuus myös epäsuorien opetusmallien tuntemisessa ja niiden hallinnan harjoittelussa oli korkeampi.

Kun opiskelijavalintaa varten kehitetään erilaisia testipatteristoja ja annetaan nimellisiä painokertoimia joillekin sen osille, on pyrkimyksenä prediktorien ja kriteerien välisen yhteiskorrelaation saaminen mahdollisimman korkeaksi. Käytännössä regressiokertoimia ei kuitenkaan ole käytetty sellaisenaan painokertoimina. Siihen on useita syitä. Toiset ovat puhtaasti teknisiä liittyen muuttujien lineaarisuuteen/ei-lineaarisuuteen. Myös otantasattuma aiheuttaa validiteettikertoimeen virheitä, kun painotettua summaa käytetään useassa koehenkilöjoukossa. Toisten syiden voidaan katsoa olevan pikemminkin sisällöllisiä: eräiden piirteiden voidaan katsoa olevan niin keskeisiä opinnoissa ja ammatissa menestymisen kannalta etteivät ne ole kompensoitavissa. Saadut tulokset antavat aihetta pohdiskeluun.

Opetus, "art of teaching", kuten Flanders (1987, 20; 1970, 270) sanoo, on moniulotteinen ilmiö, ja "jokainen opetusmalli sisältää sekä affektisen, että kognitiivisen komponentin, ja nämä molemmat on otettava huomioon jotta voitaisiin ymmärtää, mitä luokkahuoneessa tapahtuu". Tämä näkyi opetuksen ennustettavuudesta saaduissa tuloksissa selvästi. Hyvästä opettajasta voidaan sanoa: "hän oli taiteilija, mutta myös hyvä opettaja" – kuten tunnetusta opettajastani opetusneuvos Hilma Jalkasesta aikoinaan sanottiin. Samoin voidaan myös sanoa kurssilla menestyneestä opiskelijasta: "hän oli hyvä opettaja, mutta myös taiteilija". Tätä juuri epäsuoran opetuksen mallien tuottaminen ja joustavan opetuskäyttäytymisen toteuttaminen edellyttää. Flanders kuvaa myös joustavaa opetusta käytettäessä erilaisia opetusmalleja "tilanteiksi, joissa tapahtumaa katsotaan erilaisten silmälasien lävitse". Saadut tulokset antavat aihetta tutkia opetustuokiokoetta laajemmin, mm. seurantatutkimuksissa, joissa hyvin ja heikommin kokeessa menestyneiden myöhempää opinto- ja ammattimenestystä käytetään ennustettavuuden arvioinnissa kriteerinä, samoin kurssin mitattua ja arvioitua opetuskäyttäytymistä.

Opiskelijoiden kurssin evaluointitutkimuksen aineisto kerättiin käyttäen kyselomaketta, jossa oli 58 osiota ja Likert-tyyppinen 1-5 asteikko. Mittarin luotettavuus oli arvioitu aikaisemmassa vaiheessa (Heinilä 1977b). Aineisto koostui kuuden kohderyhmän mies- ja naisopiskelijoista (n = 283). *Saatujen tulosten perusteella voitiin todeta, että koulutusohjelma, opetuspaketti, didaktisen observoinnin ja pienoisopetuksen kurssi (1974 - 1991) osoittautui kokonaisuutta ajatellen toimivaksi ja opiskelijat kokivat sen varsin antoisana ja hyödyllisenä.* Yhteenvetona voitiin tehdä johtopäätös siitä, että myös nämä tulokset tukivat väittämää tausta-, panos-, prosessi-, tuotosyhteyksistä ja arviointituloksen ennustettavuudesta. Kahden eri kursseista yhdistetyn kohderyhmän - ennen ja jälkeen tutkinnon uudistuksen 1978-kurssin suorittaneiden opiskelijoiden arvioinnit poikkesivat toisistaan tilastollisesti merkittävästi kuudella seitsemästä arviointiulottuvuudesta. *Kuitenkaan koulutusohjelman sisällön ja palautejärjestelmän käyttökelpoisuuden* osalta merkittäviä arviointieroja ei esiintynyt, ei kurssiryhmien enempää kuin eri sukupuolenkaan osalta. Tämänkaltainen arvioinnin pysyvyys tuki osaltaan kurssin sisäisen validiteetin asteen arviointia. Sen sijaan mm. pääulottuvuudella, *"opetusohjelma koulutusohjelmassa", voitiin todeta eri kohderyh-*

mien arviointien välillä tilastollisesti merkitsevää eroavuutta. Jälkimmäisen ryhmän opiskelijat kokivat mm. "tarpeetonta" päällekkäisyyttä kurssin ja tiedekunnan muiden koulutusohjelmien osalta. Ilmeisesti aikaisemmin 1970-luvulla ainutlaatuisen ohjelman ainesta oli myös muissa koulutusohjelmissa lisääntyvässä määrin, mikä sinänsä lienee positiivinen ilmiö. Olihan juuri tässä opintoyksikössä jo toteutettu opettajankoulutusohjelman uudistuksessa (vuoden 1973 opettajankoulutustoimikunnan mietintö, Komiteamietintö 1975:75) sekä ohjelman sisällön tarkistuksen yhteydessä vuonna 1978 esiintuotuja näkökohtia ja tavoitteita. *Suurimpana ongelmana koettiin opiskelijoiden taholta käytännön opetustilanteisiin varatun ajan riittämättömyys.* Tämä oli odotettua, koska tutkinnon uudistuksen yhteydessä pienoisopetuksen kurssin harjoituksiin varattua aikaa oli lyhennetty 30 % entisestä. Tämä tulos tukee tiedekunnan aikaisemmin tehdyn opiskelijoiden koko opinto-ohjelman arvioinnista saatuja tuloksia (Rantakari & Tiainen 1983). Mies- ja naisopiskelijoiden arvioinneissa esiintyi myös eroja. Yleensä naisopiskelijat olivat kriittisempiä arvioinneissaan kuin miesopiskelijat, jotka mm. ilmoittivat tiedostaneensa tavoitteet heti kurssin alussa selkeästi, kun taas naiset olivat arvioinneissaan varovaisempia. Kokonaisuudessaan voitiin todeta, että opiskelijoilla arviointi täytti tilastollisen luotettavuuden vaatimukset, mitä valikoivan erotteluanalyysin perusteella saadut tulokset osoittivat: opiskelijat voitiin luokittaa erittäin suurella varmuudella: 96 % mies-, 93 % nais- ja 77 % koko otoksesta, omiin yhdistettyihin (ennen ja jälkeen tutkinnon uudistusta) kriteeriryhmiinsä (sisä- tai ulkoryhmään) arviointimuuttujien perusteella. *Nämä tulokset tukivat osaltaan olettamusta tietyn koulutusohjelman ja ympäristötekijöiden välisestä riippuvuudesta sekä koko koulutusjärjestelmästä ja siinä tehtävistä muutoksista.* Ilmeisesti opetuksen arvioinnin tulisi olla jatkuvaa ja laaja-alaista, jotta tavoitteiden saavuttamista voitaisiin arvioida ja tarvittaessa tehdä korjauksia ja muutoksia koulutusohjelmaan. Tällöin oppilaiden antama palaute on ensiarvoisessa asemassa.

Tulokset ovat osoittaneet jatkuvan pohdiskelun, tutkimuksen ja tulosten hyödyntämisen tarpeellisuuden haluttaessa kehittää opetus-oppimisolosuhteita, mm. opetusharjoittelun osalta. Opiskelijoiden tavoitteisuus määräytyy ilmeisesti suurelta osin paitsi opiskelijan luonteenominaisuuksien, myös heidän asenteittensa ja ohjelman sisällön perusteella (ks. mm. Martin et al. 2001 ja Telama et al. 1988). Tällöin motivaatioilmaston säätelyyn liittyvät tekijät, jotka ovat tämän teoriaan ja tutkimukseen perustuvan opinto-ohjelman suunnittelun lähtökohta, ovat edelleen varsin ajankohtaisia.

Totean lopuksi, että opetuskäyttäytymisen yleisiä lainalaisuuksia koskevan perustutkimuksen ohella tarvitaan selvityksiä siitä, missä määrin opiskelijat kykenevät käyttämään hyväkseen saamaansa koulutusta ja miten erilaiset yksilöt kehittyvät koulutuksen vaikutuksesta. Ilmeisesti tutkimuskohteena ollut koulutusohjelma, samoin kuin siihen liittyvä pitkän tähtäyksen tutkimusprojekti, on ollut varsin hyödyllinen: sitä todistavat osaltaan kurssin suorittaneiden ja tämän tutkimusprojektin tutkimusapulaisena toimineitten lukuisat opinnäytetyöt (pro-gradu-, laudatur-, väitöskirja- ja projektityöt) sekä heidän osallistumisensa laajaan tiedekunnan tuottamaan koulun liikuntatuntien sisältötutkimusprojektiin (ks. Varstala et al. 1983, Varstala 1996), joissa on käytetty

mm. systemaattista observointimenetelmää ja 6 sekunnin otantayksikköä aineiston hankinnassa koulun liikuntatunneilta (n = 406) opettajan (n = 248) ja oppilaiden prosessikäyttäytymisen kuvaamiseksi ja selittämiseksi.

Mainittakoon lisäksi, että observointiin ja teoriaperusteiselle evaluoinnille rakennettu koulutusohjelma on laajemminkin tunnettu; sitä on käytetty Jyväskylän yliopiston liikunnanopettajien täydennyskoulutuksessa 1970- ja 1980-luvuilla, samoin kurssia on esitetty pohjoismaisten liikunnanopettajakorkeakoulujen järjestämillä kursseilla Suomessa vuosina 1974 ja 1980 sekä Tanskassa pidetyillä Sønderborg –kursseilla vuosina 1973 ja 1975. Erittäin käyttökelpoiseksi kurssi arvioitiin mm. Brasiliassa, missä se toteutettiin Opetusministeriön pyynnöstä Rio de Janeiron yliopistossa maan korkeakoulujen opettajille (n = 40) vuonna 1978 (60 t). Koulutusohjelmaan liittyvän tutkimuksen tuloksia on niinikään raportoitu seitsemällä eri kielellä (englanti, espanja, portugali, ranska, ruotsi, saksa, suomi). Tutkimuskohteena ollut kurssi kuului liikunnanopettajien opinto-ohjelmaan esitetyssä muodossa vuoteen 1991 saakka. Vuodesta 1995 alkaen systemaattisen observoinnin ja pienoisopetuksen kurssia on toteutettu liikuntakasvatuksen laitoksella koulutusohjelmassa uudistetussa muodossa (Heikinaro-Johansson & Varstala 2000). On siis ilmeistä, että vuorovaikutusprosessin analyysijärjestelmien ja systemaattisen observoinnin tuntemus tarjoaa lisääntyvässä määrin edellytyksiä opetusharjoitteluohjelmien kehittämiselle siten, että ne voivat palvella entistä tehokkaammin korkeakoulun liikunnan opetukselle asettamien ammatillisten vaatimusten saavuttamista.

17 REFERENCES

Akkanen, O. 1979. Use of different teaching patterns based on the analysis of verbal behavior and collective movement activity/passivity of pupils in P.E. classes of junior comprehensive schoolteachers. In T. Tammivuori (ed.) Evaluation: International Congress of Physical Education, July 1976, University of Jyväskylä. Congress Proceedings of the Finnish Society for Research in Sport and Physical Education, Helsinki, 64, 89-95.

Allen, D. W. & Clark, R. J. Jr. 1967. Microteaching: Its rationale. High School Journal, 51(2), 75-79.

Allen, D. W. & Eve, A. W. 1968. Microteaching. Theory into practice, 7(5), 181-185.

Allen, D. W., Fortune, J. C. & Cooper, J. M. 1967. The Stanford summer microteaching clinic, 1965. In C. Brusling, (ed.) 1974 Microteaching - a concept in development. Stockholm: Almqvist & Wicksel International.

Allen, D. W. & Ryan, K. 1969. Microteaching. Reading, MA: Addison-Wesley.

Amidon, E. & Flanders, N. A. 1967a. The effects of direct and indirect teacher influence on dependent-prone students learning geometry. In E. J. Amidon & J. B. Hough (eds.) Interaction analysis theory, research and application. Reading, MA: Addison-Wesley, 210-216.

Amidon, E. J. & Flanders, N. A. 1967b. The role of the teacher in the classroom. Interaction analysis as a feedback system. In E. J. Amidon & J. B. Hough (eds.) Interaction analysis theory, research and application. Reading, MA: Addison-Wesley, 121-149.

Amidon, E. J. & Hough, J. B. 1967. Interaction analysis theory, research and application. Reading, MA: Addison-Wesley, 121-149.

Amidon, E. J. & Simon, A. 1965, February. Implications for teacher education in interaction research in student teaching. Paper presented at the American Education Research Association, Chicago, IL. (Eric Document Reproduction Service No. Ed 012 695)

Anderson, H. H. 1939. The measurement of domination and of socially integrative behavior in teachers' contacts with children. Child Development, 10, 73-89.

Anderson, W. G. 1971. Descriptive analytic research on teaching: Educational change in the teaching of physical education. Quest, 15, 1-8.

Anderson, W. G. & Barrett, G. T. (eds.) 1978. What's going in gym: Descriptive studies of physical education classes. Motor skills: Theory into Practice, Monograph 1.

Asetus liikuntatieteellisistä tutkinnoista no 299. Helsinki: Ministry of Education. 21.4. 1978.

Bain, L. L. 1976. An instrument for identifying implicit value in physical education programs. Research Quarterly, 47 (3), 307-315.

Bain, L. L. 1990. Physical education teacher education. In W.R. Houston (ed.) Handbook of research on teacher education a project of the association of teacher educators. NY: Macmillan, 758-781.

Bales, R. F. 1951. Interaction process analysis a method for the study of small groups. Cambridge, MA: Addison-Wesley.

Bales, R. F. & Strodtbeck, F. L. 1967. Phases in group problem solving. In E. J. Amidon & J. B. Hough (eds.) Interaction analysis theory, research and application. Reading, MA: Addison-Wesley, 89-102. (Reprinted from The Journal of Abnormal and Social Psychology, 1951, 46, 485-495).

Bandura, A. 1969. Principles of behavior modification. NY: Holt, Rinehart & Winston.

Barrett, K. R. 1969. A procedure for systematically describing teacher-student behavior in primary physical education lessons implementing the concept of movement education. Unpublished Doctoral dissertation, University of Wisconsin, Dissertation Abstracts no 70-3470.

Barrett, K. R. 1971. The structure of movement tasks: A means for gaining insight into the nature of problem-solving techniques. Quest, 15, 22-31.

Barrett, K. R. 1979. Observation for teaching and coaching. Journal of Health Physical Education and Recreation, 50 (1), 23-25.

Barrett, K. R. 1983. A hypothetical model of observing as a teaching skill. Journal of Teaching Physical Education, 3 (1), 22-31.

Barrette, G. T. 1977. A descriptive analysis of teacher behaviour in physical education classes. Columbia University Teachers College, New York. Doctoral dissertation, University Microfilms. Ann Arbor, Michigan 1978. Dissertation abstracts.

Barrette, G. T. 1996. Physical education design and research: studing curricular intents and outcomes. In G. Doll-Tepper and W.D. Brettschneider (eds.) Physical Education and Sport Changes and Challenges. Aachen: Mayer & Mayer Verlag, 142-154.

Bellack, A. A., Kiebard, H. M., Hyman, R. T. & Smith, F. L., Jr. 1966. The language of the classroom. NY: Teachers' College Press, Columbia University.

Biddle, B. J. 1967. Methods and concepts in classroom research. Review of Educational Research, 37, 337-357.

Binet, A. 1918. Les idées modernes sur les enfants (Aikamme ajatuksia lapsista). Helsinki: Otava.

Birkin, T. A. 1971. Toward a model of instructional process. In I. Westbury & A. Bellack (eds.) Research into classroom processes: Recent developments and next steps. NY: Teachers College Press, 119-137.

Bloom, B. S. 1979. Characteristiques individuelles at apprentssages Scolaires. Bruxlles: Ed. Labor

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H. & Krathwohl, D. R. (eds.) 1956. Taxonomy of educational objectives: The classification of education goals. Handbook I: Cognitive domain. NY: David McKay.

Bookhout, E. C. 1967. Teaching behavior in relation to social-emotional climate of physical education classes. Research Quarterly, 38 (3), 336-347.

Borg, W. R., Kelley, M. L., Langer, P. & Gall, M. 1970. The Minicourse: A Microteaching approach to teacher education. CA: Macmillan.

Borgatta, E. F. & Bales, R. F. 1953. Notes on research and teaching: The consistency of subject behaviour and the reliability of scoring in interaction process analysis. American Sociological Review, 18, 566-569.

Borys, A. H. 1986b. Development of a training procedure to increase pupil motor engagement time (MET). In M. Piéron & G. Graham (eds.) The 1984 Olympic Scientific Congress Proceedings. Vol. 6, Sport Pedagogy, Champaign, IL: Human Kinetics. 19-25.

Brusling, C. 1974. Microteaching: A concept in development. Stockholm: Almqvist & Wicksel International.

Campbell, D. & Fiske, D. W. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Carreiro da Costa, F. 1993. Teaching teachers: aims, methodes and contents. In J. Mester (ed.) 2nd European forum "Sport Sciences in Europe 1993" - current and future perspectives – German Sport University Cologne, September 8 – 12, 1983 conference proceedings European Network of Sport Sciences in Higher Education. Sport, leisure and physical education trends and development vol. 1, Aachen Mayer & Mayer, 484-505.

Carreiro da Costa, F., Carvalho, L., Pestana, C., Diniz, J. & Piéron, M. 1995. Physical education and sports first and fifth years students' expectations of Future work activities. In C. Paré (ed.) Better teaching in physical education? Think about it! Proceedings of the international seminar on "Training of teachers in reflective practice of physical education", Canada, Trois-Rivièrs Québec, 223-235.

Carreiro da Costa, F. & Piéron, M. 1990. Teaching learning variables related to students success an experimental teaching unit. In R. Telama, L. Laakso, M. Piéron, I. Ruoppila, V. Vihko (eds.) Physical education and life-long physical activity of the Jyväskylä Sport Congress. Jyväskylä, The Foundation for Promotion of Physical Culture and Health, 73, 304-316.

Cheffers, J. T. F. 1973. The validation of an instrument designed to expand the Flanders System of Interaction Analysis to describe non-verbal interaction, different varieties of teacher behaviour and pupil response, Temple University. Doctoral dissertation. University Microfilms 1978. Dissertation abstract 1973, No 73-23327.

Cheffers, J. T. F. 1977. Observing teaching systematically. Quest, 28, 17-28.

Cheffers, J. T. F. 1978. Systematic observation in teaching. Towards a science of teaching: Teaching analysis. In M. Piéron (ed.) University of Liège: AIESEP, vol. 2, 7-30.

Cheffers, J.T.F. 1990. Long-term research. In M. Piéron J.T.F. Cheffers & G.T. Barrette(eds.) Introduction to Sport Pedagogy: research and its applications (Reports of seminars held in Brazil. University of Liége: Boston University; Adelphi University, Garden City, 6.

Cheffers, J. T. F. & Mancini, V. 1978. Teacher-student interaction. In W. G. Anderson & G. T. Barrette(eds.) What's going in gym: Descriptive studies of physical education classes. Monograph 1, Motor skills: theory into practice, 39-50.

Clarce, S. C. T. 1971. Designs for programs of teacher education. In B. O. Smith (ed.) Research in teacher education. A symposium. NJ: Prentice Hall, INC., Englewood cliffs. N. J. 119-157.

Cloes, M., Hilbert, J. M. & Piéron, M. 1995. Effects of an observation training program on feedback, study of several cases. In C. Paré (ed.) Better teaching in physical education? Think about it! Proceedings of the International Seminar on "training on teachers in reflective practice in physical education". Trois-Rivièrs, Québec, Canada, July 1993. Universite du Québec à Trois-Rivièrs, 249-266.

Cohen, J. A. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cooley, W. C. & Lohnes, P. R. 1971. Multivariate data analysis. New York: Wiley.

Cooley, W. C. & Lohnes, P. R. 1976. Evaluation research in education: Theory, Principles and Practice. New York: Wiley.

Cronbach, L. J. 1971. Test validation. In R. L. Thorndike (ed.) Educational measurement. Washington, D.C.: American Council on Education, 443-507.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. 1972. The dependability of behavioral measurement: Generalizability of scores and profiles. New York: Wiley.

Cronbach, L. J. & Meehl, P. E. 1955. Construct validity in psychological tests. Psychological Bulletin, 7, 281-302.

Currence, J. W. 1977. Applied Behavior analysis Training Model for preservice teachers. The Ohio State University, unpublished Doctoral dissertation. Dissertation Abstract 1977, no 77-24, 615.

Darst, P., Mancini, V. & Zakrajesek, D. 1983. Use of Interaction analysis system. In P. Darst, V. Mancini & D. Zakrajsek (eds.) Systematic Observation Instrumentation for Physical Education. West Point: Leisure Press. 12-28.

Darwin, J. H. 1959. Note on the comparison of several realizations of the Markoff Chain, Biometrika, 46, 412-419.

Dillon, W. R. & Goldstein, M. 1984. Multivariate analysis, Methods and Applications. New York: Wiley & Sons.

Dodds, P. & Rife, F. 1983. Time to learn in Physical Education: History, completed research and potential future for Academic Learning Time, in Physical Education, Monograph 1, 42-47.

Doolitte, S. A., Dodds, P. & Placok, J. H. 1993. Persistante of Belifs about teaching during formal training of preservice teachers. Journal of teaching in physical education. Human Kinetics Publishers, INC. 12. 355-365.

Dougherty, N. J. 1970. A comparison of the effect of command, task and individual program styles of teaching in the development of physical education fitness and motor skills. Temple University. Unpublished Doctoral dissertation, University Microfilms 1971. Dissertation abstract No. 71-10, 813. (5821-22A)

Dougherty, N. J. 1971. A plan for the analysis of teacher-pupil interaction in physical education classes. Quest, 15, 39-49.

Dougherty, N. 1983. Adaptation of the Flanders system of interaction analysis. In: P. Darst, V. Mancini & D. Zakkraisek (eds.) Systematic observation in strumentation for physical education. West Point: Leisure Press 129-133.

Dunkin, M. J. & Biddle, B. J. 1974. The study of teaching. NY: Holt, Rinehart & Winston.

Dunkin, M. J. 1976. Towards taking some of the fun out of the technical skills approach to teacher effectiveness. South Pacific Journal Teacher Education 4 (2), 131-39.

Dunkin, M. J. 1987. Technical skills of teaching. In M. J. Dunkin (ed.) The international encyclopedia of research teaching and teacher education. Oxford: Pergamon Press, 703-706.

Dunkin, M. J. 1987 (ed.). The international encyclopedia of teaching and teacher education. Oxford: Pergamon Press.

Ebbs, P. A. 1975. The relationship among selected characteristics of cooperating teachers and the attitude change of student teachers. The Temple University. Unpublished Doctoral dissertation. Dissertation abstracts 1976. No 75-28, 272. (4415-A)

Educational development in Finland 73-75, 1975. Reference Publication 7. Helsinki: Ministry of education.

Emmer, E. T. 1972. Direct observation of classroom behavior. International Review of Education, 18(4), 473-490.

Evertson, C. M. & Green, J. L. 1986. Observation as inquiry and method. In M. C. Wittrock (ed.) Handbook of research on teaching (3rd ed.) NY: McMillan, 162-213.

Feingold, R. S. 1972. The evaluation of the teacher education programs in physical education. Quest, 18(June), 33-40.

Feingold, R. S. & Barrette, G.T. 1988. Modernization of preservice teacher preparation program: Critical issue. In: Reader H. And Hanke U. (eds.) The physical education teacher and coach today. Sportlehrer und Trainer heute. Band 2, Vol. 2, Bundesinstitut für Sportwissenschaft, Köln, 30-35.

Finske, S. M. J. 1967. The Effect of Feedback Trough Interaction Analysis on the development of Flexibility in student teachers. University of Michigan. Unpublished Doctoral dissertation. Dissertation abstracts 1967, no 67-15, 621.

Fishman, S. & Anderson, W. 1971. Developing a system for describing teaching: Educational change in the teaching of physical education. Quest, 15, 9-16.

Flanders, N. A. 1965. Teacher influence, pupil attitudes and achievement (HEW, Office of Education Cooperative Research, Monograph 12. Washington: D.C.: U.S. Government Printing Office.

Flanders, N. A. 1966. Subscripting interaction analysis categories: A 22 category system. Ann Arbor: University of Michigan.

Flanders, N. A. 1967a. The problems of observer training and reliability. In E. J. Amidon & J. B. Hough (eds.) Interaction analysis: Theory, research and application. Reading, MA: Addison-Wesley, 158-166.

Flanders, N. A. 1967b. Teacher influence in the classroom. In E. J. Amidon & J. B. Hough (eds.) Interaction analysis: Theory, research and application. Reading, MA: Addison-Wesley, 103-116.

Flanders, N. A. 1970. Analyzing teaching behavior. Reading, MA: Addison-Wesley.

Flanders, N. A. 1987. Human interaction models. In M. J. Dunkin (ed.) The international encyclopedia of teaching and teacher education. Oxford: Pergamon Press, 10-28.

Flanders, N. A. 1987. Flexibility. In M. J. Dunkin (ed.) The international encyclopedia of teaching and teacher education. Oxford: Pergamon Press, 462-466.

Flanders, N. A. & Nuthall, G. (eds.) 1972. International review of education, 18.4, Special number: The classroom behavior of teachers. Hamburg: UNESCO Institute of Education.

Flanders, N. A. & Simon, A. 1970. Teaching effectiveness: A review of research 1960-'66. In R. L. Ebel (ed.) Encyclopedia of educational research. Chicago: Rand McNally.

Foa, U. G 1965. New development in facet design and analysis. Psychological Review, 72, 262-274.

Gage, N. L. 1963b. Paradigms for research on teaching. In N. L. Gage (ed.) Handbook of research on teaching. Chicago: Rand McNally, 94-141.

Gage, N. L. 1969. Teaching Methods. In R. L. Ebel (ed.) Encyclopedia of Educational Research. (4th ed.) New York: Macmillan, 1446-1458.

Gage, N. L. 1972. Teacher effectiveness and teacher education. Palo Alto, Cal.: Pacific Books.

Gage, N. L. 1978. The scientific basis of the art of teaching. (4th ed.) NY: Teachers College Press, 14-17, 44.

Gage, N. L. & Berliner, D. C. 1979. Educational psychology. (2nd ed.) Chicago: Rand McNally, 635-703.

Galloway, C. M. 1962. An exploratory study of observational procedures for determining teacher nonverbal communication. University of Florida, Doctoral dissertation, University Microfilms. Dissertation abstract No. 62, 6529.

Galloway, C. M. 1966. Teacher nonverbal communication. Educational Leadership, 24, 55-63.

Galloway, C. M. 1968. Theory into practice. Nonverbal Communication, 7 (5), 172-175.

Galloway, C. M. 1970. Teaching in communication, nonverbal language in the classroom (Bulletin No. 29). Association for Student Teaching, 14-6.

Galloway, C. M. 1971. Teaching is more than words. Quest, 15, 67-71.

Garrett, F. D. 1973. Feedback and Flanders interaction analysis related to change in the indirect teaching behavior of student teachers, Northern Illinois University, unpublished Doctoral dissertation, Dissertation Abstracts 1973, No 73-20544. 1161-A.

Gasson, I. S. H. 1971. The development of an observational instrument to record selected teacher-pupil behaviors in primary school physical education. The Ohio State University. Doctoral dissertation. Ann Arbor, MI: University Microfilms. Dissertation Abstract No. 71-27, 473. (1901-A)

Gorman, A. H. 1969. Teachers and learners: The interactive process of education. Boston: Allyn and Bacon.

Greenwood, G.E. & Ramagli, H. J. Jr. 1980. Alternatives to student ratings of college teaching. Journal of Higher Education, 6, 673-684.

Guttman, L. P. 1954. A new approach to factor analysis: The radex. In P. F. Lazarsfeld (ed.) Mathematical thinking in social sciences. Glencoe, IL: The Free Press.

Hanke, U. 1976. The evaluation of teaching skills. FIEP Bulletin, 46 (3), 66-72.

Hanke, U. 1979. The importance of evaluation in modelling and feedback for the acquisition of teaching-skills. In T. Tammivuori (ed.) Evaluation international congress of physical education, July 1976 (Reports of the Finnish Society for Research in Sport and Physical Education). Helsinki: Finnish Society for Research in Sport and Physical Education, 64, 74-80.

Hanke, U. 1980. Training des Lehrerverhaltens von Sportstudenten. Ein Vergleich zweier Trainingsverfahren auf der Basis des Microteaching (Training of the teacher behavior of sport students). Inaugural Dissertatian, Sozial- und Verhaltenswissenschaftlichen Fakultät der Ruprecht-Karls-Universität zu Heidelberg. Heidelberg: Udo Hanke.

Hanke, U. 1986. Methodological considerations on physical education teacher training. In C. Paré, M. Lirette and M. Piéron (eds.) Research methodology in teaching physical education and sports: international seminar. Methodology Departement des Sciences de l'activite physique. Universite du Québec à Trois-Rivières. Trois-Rivières, Québec, Canada, 123-140.

Hanke, U. & Treutlein, G. 1983. What P. E. teachers think: methods for the investigation of P. E. teacher cognitions in teaching process. In R. Telama, V. Varstala, J. Tiainen, L. Laakso & T. Haajanen (eds.) Research in School Physical Education. Reports of Physical Culture and Health 38. Jyväskylä: Foundation for promotion on physical Culture and Health, 31-37.

Harrington, W. 1974. A Study of Feedback Diversity in Teaching Physical Education. University of Wisconsin-Madison, unpublised Doctoral dissertation, Xerox University Microfilms, 1976. Ann Arbor, MI: Dissertation abstracts 1974, No 74-30, 105.

Heikinaro-Johansson, P. & Varstala, V. 2000. Developing the teaching skills of physical education students through self-evaluation. In F. Carreiro da Costa, J. A. Diniz, L. M. Carvalho & M. S. Onofre (eds.) Research on teaching and research on teacher education: Proceedings of the Lisbon AIESEP International Seminar. Portugal: Cruz Quebrada FMH, 221-225.

Heinilä, L. 1970. Opettajan ja oppilaiden välisistä vuorovaikutussuhteista liikunnan opetustilanteissa. (About teacher-pupil interaction in physical education classes.) Reports of the Finnish Society for Research in Sports and Physical Education, 22. Helsinki: The Finnish Society for Research in Sport and Physical Education, 80-94.

Heinilä, L. 1971. Liikunnan opetustapahtuma sosiaalisena vuorovaikutusprosessina (Teaching of physical education as a process of social interaction). University of Jyväskylä, Finland. Unpublished master's thesis.

Heinilä, L. 1974. Developing a system for describing teacher-pupil interaction in physical education classes. Paper presented at FIEP scientific congress Gdansk 27-31. May 1974. In T. Bober. and G. Mlodzikowski (eds.). Education physique des enfants avant l'Epoque de la Puberte. Edition Scientifiques de Pologne, Warsaw, Monographie no 12, Gdansk 1976, 218-223, and FIEP Bulletin 1974, 44(4), 16-20 (Eng.), 59-62 (French).

Heinilä, L. 1976. Process objectivity of coding in a system (PEIAC/LH-75) developed for describing teacher-pupil interaction in physical education classes. In T. Haajanen and M. Veistola (eds.) Research in Physical Culture in Finland, Policy in Physical Culture Research Work, Abstracts IV 1976. Reports of the Finnish Society for Research in Sports and Physical Education, 1977, 55 and 49. Helsinki: Finnish Society for Research in Sports and Physical Education, 66, 22-23.

Heinilä, L. 1977a. Analysing systems in the evaluation of the teacher-pupil interaction process in physical education classes. In Tammivuori (ed.), Evaluation: International Congress of Physical Education, July 1976, University of Jyväskylä. Congress proceedings of the Finnish Society for Research of Physical Education and Sport no. 64. Helsinki, 1979, 37-58. FIEP Bulletin 1977, 47(1), 20-34 (Eng.), 47(1) 13-25 (French). FIEP Bulletin 1978 48(3), 4-23 (Portug.). Methode d' evaluation du processus d'enseignemet en education physique, FFGEV-Gymnastique. Volontaire 1, 1977, 24-33 (French).

Heinilä, L. 1977b. Application of interaction analysis to the teacher education in physical education. Paper precented at the International AIESEP-FIEP Congress of Physical education and Sports, Madrid June, 1977. Research reports from the Departement of Physical Education, University of Jyväskylä, 15. (1979) and Research Bi-Annual for movement. Manhattan-State India 13 (2) 1997, 16-56.

Heinilä, L. 1980. Developing a system (PEIAC/LH-75) for describing teacher-pupil interaction in physical education classes: Objectivity and content validity of coding. In G. Schilling & W. Bauer (eds.) Audiovisual Means in Sport. Basel: Birkhaus Verlag, 361-370.

Heinilä, L. 1983. Developing a system (PEIAC/LH-75) for describing teacher-pupil interaction in physical education classes: Construct validity and sensitivity. In R. Telama, V. Varstala, J. Tiainen, L. Laakso & T. Haajanen (eds.) Research in school physical education. AIESEP congress 1982 Jyväskylä. Finland. Reports of the Foundation for Promotion of Physical Culture and Health, 38, 124-132.

Heinilä, L. 1987. The Development, validitation and application to teacher training of a system (PEIAC/LH-75) designed to expand the Flanders system of interaction analysis for describing teacher-pupil interaction process in physical education classes, unpublished lic. thesis University of Jyväskylä; 20.10.1987.

Heinilä, L. 1988 Selecting students for physical education teacher education programmes. FIEP. Bulletin 58, 2/3 1988, 29-42.

Heinilä, L. 1990. Validation of an observation system in physical education: A multivariate approach. Paper presented at the International AIESEP Congress, Trois-Riviers, Québec, Canada. 1987. In M. Lirette, C. Paré, J. Dessureault & M. Piéron (eds.) Intervention en Éducation Physique et en Entrainement, Bilan et Perspectives. Physical Education and Coaching, Present State and Outlook for the Future. Québec: Presses de l'Université du Québec, 28-40.

Heinilä, L. 1992a. Prediction of success in student teaching from students selection variables, rated and measured teaching behavior and students' attitudes. Research report presented at the Olympic Scientific Congress 14.-19.6.1992 in Malaga, Spain. Actas Congreso Cientifico Olympico 1992. Pedagogia y Education Fisica Comparada. Serie Deporte y Documentation Instituto Andaluz Del Deporte no. 24, 1995, vol. III, 54-62. Also in references C. Paré (ed.) Better teaching in Physical Education? Think about it! (1995) Canada, Trois-Rivièrs: University of Québec, 291.

Heinilä, L. 1992b. Relationship between student teachers expectations concerning "ideal" P.E. teacher characteristics and their own teaching behavior and success in a microteaching course. Paper presented at the FIEP World Congress, in Nabeul, Tunisia.

Hendry, L. B. 1969. A Personality Study of highly successful and "ideal" swimming coaches. Research quartely, 1969, 40, 299-305.

Hendry, L. B. 1978. Conflicts in the curriculum: an example from physical education. Educational Research 2 (3), 174-180.

Higher Education and research in Finland. Reference. Publications 6. Helsinki: Ministry of Education. 1973.

Honigman, F. K 1970. Multidimensional analysis of classroom interaction (MACI): The Honigman system of interaction analysis. In A. Simon & E. G. Boyer (eds.) Mirrors for behavior: An anthology of classroom observation instruments (Vol. II). Philadelphia: Research for Better Schools, Inc.

Hough, J. & Ober, R. 1967. The effect of training in interaction analysis on the verbal teaching behavior of pre-service teachers. In E. J. Amidon & J. B. Hough (eds.) Interaction analysis: Theory, research and application. Reading, MA: Addison-Wesley, 329-345.

Hupé, A. 1995. The seminar: A conclusion. In C. Paré (ed.) Better teachning in physical education? Think about it! Proceedings of the international seminar on "training of teachers in reflective practice in physical education" Canada, University du Quebéc, Trois-Riviérs, 277-289.

Hytönen, J. & Komulainen, E. 1971. Opettajan kyvystä noudattaa annettua opetustyyliä. (About teacher's ability to follow of given teaching style.) Kasvatustieteen opiskelijoille suoritettu kokeellinen tutkimus. Kasvatus 2, 98-111.

Hytönen, J. 1973. Opettajakokelaiden opetusharjoitus käyttäytymisen selittämisestä eräiden persoonallisuuspiirteiden avulla. (About realtionships of students preservice teaching behavior and some personality characteristics variables.) Helsingin yliopiston kasvatustieteen laitos. Väitöskirja. Tutkimuksia no 26.

Itälä, J. 1969. Koulutussuunnittelu ja koulusuunnittelu (Planning of schooling and school planning). In J. Itälä (ed.) Koulusuunnittelu. Helsinki: Tammi, 11-22.

Jawett, E. & Müllan, M. R. 1972. A conceptual model for teacher education. Quest, Spring Issue, Monograph 18. 76-87.

Joyce, B. & Weil, M. 1980. Models of teaching. NJ: Englewood Cliffs, Prentice-Hall.

Joyce, B. R., Brown, C. C. & Peck, L. (eds.) 1981. Flexibility in teaching, New York: Longman.

Jyväskylän yliopiston liikuntatieteellisen tiedekunnan tutkintosääntö. (Examination requirements of the department of physical education of the Univeristy of Jyväskylä.) Ministry of Education Helsinki 31.8.1972.

Jyväskylän yliopiston liikuntatieteellisen tiedekunnan opinto-opas (Study quide department of physical education of the University of Jyväskylä) 1974-75, 1975-76, 1976-77, 1977-78, 1979-80, 1986-87.

Kane, J. E. 1968. Personality in relation to physical abilities and physique. University of London. Doctoral dissertation. Abstract. Index to theses, vol. XVIII 1967, 68, 54.

Keilty, G. C. 1975. The effect of instruction and supervision in interaction analysis on the preparation of student teachers. Boston University. Unpublished Doctoral dissertation. University Microfilms 1980 dissertation. Abstract. No 75-20, 956.

Kemper, H., Ras, J., Verschuur, R., Snel, J., Splinter, P. & Tavecchio, L. 1976. Development of an instrument for the analysis of the social-emotional teacher-pupil interaction in physical education. In proceeding of the FIEP congress 1974: Education physique des enfants avant l'époque de la puberté. Warsaw: Edition Scientifiques de Pologne, 234-239.

Kerlinger, F. N. 1973. Foundations of behavioral research (2nd ed.) 1973. London: Holt, Rinehart & Winston.

Kirk, D. 1986 "A critical pedagogy for teacher education: toward an inquiry-oriented approach" Journal of Teaching Physical Education, 5, 230-246.

Kirk, D. 1993. Curriculum work in physical: Beyonds the objective approach? Journal of Teaching in Physical Education, 12(3), 244-265.

Komiteanmietintö. 1970a. Peruskoulun opetussuunnitelmakomitean mietintö I: opetussuunnitelman perusteet (No. A4) (Report of the national commission for primary school I: basics for planning of teaching). Helsinki: Valtion painatuskeskus.

Komiteanmietintö. 1970b. Peruskoulun opetussuunnitelmakomitean mietintö II: oppiaineiden opetussuunnitelmat (No. A5) (Report of the national commission for primary school II: planning of the teaching subjects). Helsinki: Valtion painatuskeskus.

Komulainen, E. 1968. Opetustapahtuman tutkimuksesta observointi-menetelmällä (On the study of teaching by observation), University of Helsinki, Unpublished phil.lic. thesis.

Komulainen, E. 1970. Investigations into the instructional process (Vol. II): Objectivity of coding in a modified Flanders Interaction Analysis. University of Helsinki. Studies in Education 27.

Komulainen, E. 1971a. Investigation into the instructional process (Vol. III): P-technique treatment of observational data. University of Helsinki, studies in Education 28.

Komulainen, E. 1971b. Investigation into the instructional process (Vol. IV): Teaching as a stochastic process. University of Helsinki, Studies in Education 29.

Komulainen, E. 1973. Investigation into the instructional process (Vol. VIII): On the problems of variable construction from Flanders' interaction matrix with special emphasis on the stochastic nature of classroom communication. University of Helsinki, Studies in Education 34.

Komulainen, E. 1974a. Sattumakorjattujen yksimielisyyskertoimien käytöstä luokitteluun perustuvan tutkimusaineiston yhteydessä (Chance-corrected agreement coefficients applied to nominal data). University of Helsinki. Studies in Education 33.

Komulainen, E. 1974b. Opettajankoulutus ja opetusharjoittelun uudet muodot. (teacher education and the new forms of practice teaching) Suomen kasvatustieteellinen aikakauskirja kasvatus 5 (4), 227-232.

Komulainen, E. 1978. Developmental change in interaction patterns of the DPA classes. In E. Komulainen & M. Koskenniemi (eds.) DPA Helsinki Investigations II: Research on Teaching. University of Helsinki: Studies in Education, 17-28.

Komulainen, E. and Kansanen, P. (eds.) 1981. Classroom analysis, Findings Applications. DPA Helsinki Investigations III. Reserach reports of Institute of education. University of Helsinki 56.

Koskenniemi, M. 1981. Activity forms and the formal proprieties of instruction. In E. Komulainen & P. Kansanen (eds.) Classroom analysis findings, applications DPA Helsinki investigations III. Research reports of Institute of Education, University of Helsinki 56. 39-52.

Koskenniemi, M. & Hälinen, K. 1970. Didaktiikka. Helsinki: Otava.

Koskenniemi, M. & Komulainen, E. 1969. Investigation into the instructional process (Vol. I): Some Methodological Problems. Helsinki, Finland: University of Helsinki, Studies in Education 26.

Koskenniemi, M., Komulainen, E., Kansanen, P., Karma, K., Matikainen, M., Holopainen, P. & Uusikylä, K. (eds.) 1974. DPA Helsinki. System for describing Instruction and Processes. Research reports of University of Helsinki. Studies of Education 42.

Koskenniemi, M. & Komulainen, E. 1978. Research on teaching. Papers presented at the symposium arranged by the Academy of Finland for evaluation of DPA-Helsinki: project, Helsinki, Oct. 26th and 27th 1977. DPA Helsinki Investigations II ed. by Erkki Komulainen and Matti Koskenniemi 1978.

Kuhn, T. S. 1962. The structure of scientific revolutions. International Encyclopedia of Unified Science: Vol. 2, No. 2. Chicago: University of Chicago Press.

Laakso, L. 1975. The motives for career choice of student aiming at becoming Physical Education teachers. Journal of the Finnish Society for Research in Sport and Physical Education, Stadion, 12. 1975, 35-38 and in T. Haajanen & M. Veistola (eds.) 1977 Research in Physical Culture in Finland, Policy in Physical Culture Reserach Work, Abstracts IV, 1976. Reports of the Finnish Society for Research in Sports and Physical Education. No 55. Helsinki: Finnish Society for Reserach in Sport and Physical Education, 54.

Laakso, L. 1984. The Norwegian, Swedish and Finnish Student in Physical Education Teacher Training I. Theoretical background, methods and preliminary results. University of Jyväskylä, Reports from the Department of Physical Education 17.

Lamarre, G. & Nygaard, G. 1977, June. A comparison of the effects of command and guided discovery styles of teaching on college students' cognitive achievement and their attitudes toward a winter camping class. Paper presented at the meeting of the International Congress of AIESEP, Madrid.

Lawson, H. A. 1988. Occupational Socialization, Cultural Studies and the Physical Education Curriculum. Journal of teaching Physical Education, 7, 265-288.

Levin, L. 1968. Observationer av elevaktiviteter under gymnastiklektioner (Observations of student activities during physical education lessons) Göteborgs Universitet, reports from the Pedagogiska Institutionen 27.

Lewin, K. 1935. Dynamic theory of personality. NY: H. W. Wildon.

Lewin, K., Lippitt, R. & White, R. 1939. Patterns of aggressive behavior in experimentally created 'social climates'. Journal of Social Psychology, 10, 271-299.

Lippitt, R. & White, R. K. 1943. The 'social climate' of children's groups. In R. G. Barker, J. S. Kounin & H. F. Wright (eds.) Child Behavior and Development. NY: McGraw-Hill.

Locke, L. F. 1977. Research on teaching physical education: New hope for dismal science. Quest, Monograph 28, 2-16.

Locke, L. F. 1983. Research on teacher education for physical education in the U.S.A., Part II Questions and conclusions. In R. Telama, V. Varstala, J. Tiainen, L. Laakso & T. Haajanen (eds.) Research in school physical education. Reports of physical Culture and Health 38. Jyväskylä: Foundation for Promotion of Physical Culture and Health, 285-320.

Locke, L. F. 1984. Research on Teaching Teachers: Where are we now? Journal of Teaching in Physical Education, Mongraph 2, 2-16; 3-86.

Locke, L. F. 1986. Qualitative research in gym: Old problems and new answers. In C. Paré, M. Lirette and M. Piéron (eds.) Research methodology in teaching physical education and sports: international seminar. Methodology Departement des Sciences de l'activite physique. Universite du Québec à Trois-Rivières. Trois-Rivières, Québec, Canada, 35-56.

Locke, L. F. 1989. Qualitative Research as a form of scientific Inquiry in Sport and Physical Education. Research Quarterly for Exercise and Sport, v. 60, 1, 1-20.

Locke, L. F. & Dodds, P. 1981. Research on preservice teacher education for physical education in U.S.A. Paper presented in third AIESEP seminar Rio de Janeiro, july 20, 1981.

Lombardo, B. & Cheffers, J. 1983. Variability in teaching behavior and interaction in gymnasium. Journal of Teaching in Physical Education. 2, 2. 33-48.

Love, A. & Roderick, J. 1971. Teacher nonverbal communication: the development and field testing of an awareness unit. Theory Into Practice: The Challenge of Nonverbal Awareness, 4, 295-299.

Lundgren, U. P. 1972. Frame factors and teaching process. Stockholm: Almqvist & Wiksell.

Mancini, V. H. & Cheffers, J. T. F. 1983. Cheffers' Adaptation of Flanders interaction analysis system II (CAFIAS). In P. Darst, V. Mancini & D. Zakrajsek (eds.) Systematic Observation instrumentation for Physical Education. West Point: Leisure Press, 96-99.

Mancuso, J. T. 1972. The verbal and nonverbal interaction between secondary school physical education student teachers and their pupils. University of Illinois, Doctoral dissertation. Ann Arbor, Michigan: University Microfilms 1973, dissertation abstract No 73-17310, 606-A.

Martel, D. 1995. The Education to a Reflective Practice: A process of Shared Learning. In C. Paré (ed.) Better teaching in Physical Education? Think about it! Proceeding of the international seminar on training of teachers in reflective practice in Physical Education. Canada: Université du Québec a Trois Rivièrs, 88-111.

Martens, F. L. 1987. Selection of physical education students and success in student teaching. Journal of Teaching Physical Education, 6 (4), 411-424.

Martin, J. J., Hodges Kulinna, P., Eklund, R. C. & Reed, B. 2001. Determinants of teachers' intentions to teach physically active physical education classes. Journal of Teaching in Physical Education, 20, 129 – 143.

McGaw, B., Wardrop, J. L. & Bunda, M. A. 1972. Classroom observation schemes: Where are the errors? American Educational Research Journal, 9 (1), 13-27.

Medley, D. M. 1971. The language of teacher behavior: communication of the results of structures observations to teacher. The Journal of Teacher Education, 22 (2), 157-165.

Medley, D. M. 1982. Systematic observation. In H. E. Mitzel (ed.) Encyclopedia of educational research (5th edition). NY: The Free Press, 1841-1851.

Medley, D. M. 1987. Evolution of research on teaching. In Dunkin M. J. (ed.) (1987) The international encyclopedia of teaching and teacher education. Oxford: Pergamon Press, 105-113, 169.

Medley, D. M. & Mitzel, H. E. 1958. A technique for measuring classroom behavior. Journal of Educational Psychology, 49 (2), 87-92.

Medley, D. M. & Mitzel, H. E. 1963. Measuring classroom behavior by systematic observation. In N. L. Gage (ed.) Handbook of Research on Teaching. Chicago: Rand McNally, 247-328.

Melograno, V. 1971. Effects of teacher personality, teacher choice of educational objectives and teacher behavior on student achievement. Temple University. Unpublished Doctoral dissertation. University Microfilms 1972. Dissertation abstract No 72-20, 202.

Melograno, V. J. 1979. Design curriculum and learning a physical coeducational approach. Dubuque, IA: Kendall, Hunt Publishing Company, 289, 381.

Melograno, V. J. 1985. Designing the physical education curriculum a self-directed approach. Dubuque, IA: Kendall, Hunt Publishing Company.

Mitzel, H. E. 1960. Creiteria of Teacher effectiveness. In C. W. Harris (ed.) Encyclopedia of Educational Research. (3rd ed.) NY: Macmillan, 1481-86.

Mosston, M. 1966. Teaching physical education. Columbus, OH: Merrill Publishing Company.

Nixon, J. E. & Locke, L. F. 1973. Research on teaching physical education. In R. M. W. Travers (ed.) Second Handbook of Research on Teaching Chicago: Rand McNally, 1210-1242.

Nygaard, G. 1971. An Analysis of verbal interaction in physical Education classes. University of Oregon. Unpublised Doctoral Dissertation. Abstract.

Nygaard, G. 1978. Three research papers on the analysis of teaching in physical education. In M. Piéron (ed.) Towards a science of teaching physical education: Teaching analysis vol. 2. Liège: AIESEP, 53-58.

Pankratz, R. 1967. Verbal interaction patterns in the classrooms of selected physics teachers. In E. J. Amidon & J. B. Hough (eds.) Interaction analysis: Theory, Research and Application. Palo Alto, CA: Addison-Wesley, 189-209.

Paré, C. (ed.) 1995. Better teaching in Physical Education? Think about it! Proceedings of the international Seminar on "training of teachers in reflective practice of physical education". Canada: Université du Québec à Trois-Rivièrs, i-vii.

Paré, C., Lirette, M., Laurencelle L. 1986. Methodological report on research on teaching physical education and sports. In C. Paré, M. Lirette and M. Piéron (eds.) Research methodology in teaching physical education and sports: international seminar. Methodology Departement des Sciences de l'activite physique. Universite du Québec à Trois-Rivières. Trois-Rivières, Québec, Canada, 105-122.

Parsons, T. 1968. Social interaction. In D. L. Sills (ed.) International Encyclopedia of The Social Sciences vol. 7. NY: Macmillan The Free Press, 429-441.

Pedhadzur, E. J. 1982. Multiple regression in behavioral research. Explanation and prediction. NY: Holt, Rinehart and Winston.

Piéron, M. 1978. Relationships Between verbal and nonverbal behaviors in teaching physical education. In M. Piéron (ed.) Towards a science of teaching physical education: Teaching analysis vol 2. Liège: AIESEP, 69-76.

Piéron, M. 1982. Effectiveness of teaching psycho-motor task. Study in a microteaching setting. In M. Piéron & J. Cheffers (eds.) Studing the teaching in physical education. Liége. AIESEP, 79-89.

Piéron, M. 1983. Teacher and pupil behavior and the interaction process in P.E. classes. In R. Telama, V. Varstala, J. Tiainen, L. Laakso & T. Haajanen (eds.) Research in School Physical Education. Proceedings of the International Symposium on Research in School Physical Education (1982): Reports of Physical Culture and Health 38. Jyväskylä: The foundation for Promotion Physical Culture and Health, 13-30.

Piéron, M. 1984. Pedagogies des activités physiques et sportives (méthodologie et didactique). Université de Liège.

Piéron, M. 1986. Analysis of the Research Based on observation of the teaching Physical Education. In M. Piéron & G. Graham (eds.) The 1984 Olympic Scientific Congress Proceedings vol. 6. Sport Pedagogy, Champaign, IL: Human Kinetics, 193-202.

Piéron, M. 1989. Preface AIESEP Outcomes and Perspectives. AIESEP-Weltkongress von 22.-26. August 1986. In H. Rieder & U. Hanke (eds.) The physical education teacher and coach today, vol. 2. Köln, Bundesinstitut für Sportwissenschaft.

Piéron, M. 1992. Highlights on Teaching effectiveness. Research studies. University of Liège. 26.

Piéron, M. 1993. Educational Research in Physical Education. In J. Mester (ed.) 2nd European Forum "Sport Sciences in Europe 1993" - Current and Future perspectives. European Network of Sport Sciences in Higher Education. Sport, Leisure and Physical Education Trends and development vol. 1, Aachen: Mayer & Mayer Verlag, 611-641.

Piéron, M. 1994. Sport pedagogy. Highlights on research on teaching. Research on teacher preparation. Liège: University of Liège.

Piéron, M. 1996. Selected research trends in sport pedagogy in R. Lidor, E. Eldar and I. Harary (eds.) AIESEP World Congress 1995. windows to future: bridging gaps between diciplines, curriculum and instruction. Israel the Wingate Institute, I, 56-69.

Piéron, M. & Cheffers, J. 1988. Research in Sport Pedagogy. Empirical analytical perspective ICSSPE Sport Science Studies. 2. Schorndorf: Verlag Karl Hofmann.

Piéron, M., Cheffers, J. & Barrett, G. 1990. An Introduction to the Terminology of Sport Pedagogy (vocabulary used in research in teaching and coaching). Liège: International Committee of Sport Pedagogy (ICSP).

Piéron, M. & Haan, J. 1980. Pupils activities, time on task and beahaviours in high school physical education teaching. FIEP Bulletin 50 (3-4), 62-68.

Piéron, M. & Piron, J. 1981. Research de critères d'efficacité de l'einseignement d'habilites motrices. Sport, 24. 144-161.

Pitkänen, P., Komi, P. V., Nupponen, H., Rusko, H., Telama, R. & Tiainen, J. 1979. Evaluating the product of physical education. In T. Tammivuori (ed.) Evaluation: International Congress of Physical Education, July 1976. Reports of the Finnish Society for research in Sport and Physical Education 64. Helsinki. Finnish Society for Researcy, 119-136.

Placek, J. & Dodds, P. 1988. A critical study of preservice teachers' beliefs about teaching success and nonsuccess. Research quartely for Exercise and Sport. 59, 351-358.

Quilford, J. 1948. Factor analysis in a test development paradigm. Psychological Review, 55, 79-94.

Rantakari, J. & Tiainen, J. 1983. P.E. Student ratings of teaching and problems encountered during studies. In R. Telama et al (eds.) Research in School Physical Education. Reposts of Physical Culture and health 38. Jyväskylä: Foundation for Promotion of Physcial Culture and health, 336-342.

Reponen, P. 1979. Personality and teaching behavior in physical education students. In T. Tammivuori (ed.) Evaluation: International Congress of Physical Education. Jyväskylä, July 1976. Reports of the Finnish Society for Research in Sport and Physical Education, 64, 96-115.

Rogers, C. R. 1967. The facilation of Significant learning. In L. Siegel (ed.) Instruction. Some contemporary view points. San Fransisco: Chandler, 37-54.

Rogers, C. R. 1980. A way of being. Aspects, personcentered approach. Boston: Houghton Mifflin Company.

Rogers, J. E. 1969. Methode of improving the professional preparation. Teachers Research Quarterly.

Rosenshine, B. 1970. Evaluation of classroom in struction. Review of Educational Research 1970, 40, 279-300 (a).

Rosenshine, B. 1971. Teaching behaviours and student achievement. Slough, England: National Foundation for Education in England and Wales.

Rosenshine, B. 1976. Recent Research on Teaching Behaviors and Student Achievement. Journal of Teacher Education. 27 (1), 61-64

Rosenshine, B. & Furst, N. 1971. Research in teacher performance criteria. In B.O. Smith (ed.) Research in teacher Education (A Symposium). Englewood Cliffs, N.J.: Prentice Hall, 37-72.

Rosenshine, B. & Furst, N. 1973. The use of direct observation to study teaching. In R.M.W. Travers (ed.) Second Handbook of Research on Teaching. Chicago: Rand McNally, 122-183.

Rowley, G. I. 1976. The reliability of observational measures. American Educational Research Journal, 13 (1), 51-59.

Rudy, E. 1974. A comparative study of the Effectiveness of the small-group Method and command univeristy of New Mexico. Unpublished doctoral dissertation. Xerox University Microfilms International. 1976. Ann Arbor, Michigan USA, No 48106.

Safrit, M. J. 1973. Evaluation in physical education. Englewood Cliffs, NJ: Prentice-Hall.

Schéffe, H. 1959. The analysis of variance. (2nd ed. 1969) NY: Wiley, 73-77.

Schempp, P. 1985. Becoming Better teacher: Analysis of the Student experience. Journal of Teaching Physical education, 4. 158-166.

Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 19, 321-325.

Siedentop, D. 1972. Behavior analysis and teacher training. Quest, Spring Issue, Monograph 18, 26-32.

Siedentop, D. 1981. The Ohio State University Supervision reserch program summary report. Journal of Teaching in Physical Education, Introductory Issue, 30-38.

Siedentop, D. 1983. Developing teaching skills in Physical Education (2nd ed.) Palo Alto CA: Mayfield.

Siedentop, D. 1986. The modification of Teacher Behavior. In M. Piéron & G. Graham (eds.) The 1984 Olympic Scientific Congress Proceedings, vol. 6, Sport Pedagogy. Champaign, Human Kinetics. 3-18.

Siegel, S. 1956. Nonparametric statistics of the behavioral sciences. Tokyo: McGraw-Hill.

Silvennoinen, M., Laakso, L. & Turunen, J. 1991. Liikuntatieteellisen tiedekunnan valintatutkimus liikunnanopettajakoulutuksen vuosina 1978-1982 and 1986 hyväksyttyjen opiskelijoiden koulu- ja valintakoemenestys ja opintomenestys. Jyväskylän Yliopiston Liikuntakasvatuksen laitos, Tutkimukssia A 20.

Silverman, S. 1991. Research on Teaching Physical Education, Research quarterly for Exercises and Sport. American Alliance for Heath, Physical Education and Recreation and Dance, 62, 4, 352-364.

Silverman, S. & Skonie, R. 1997. Research on teaching in Physical Education: An analysis of published research. Journal of Teaching in Physical Education, 1997, 16, 300-311.

Simon, A. & Boyer, E. (eds.) 1970. Mirrors for behavior: An anthology of observation instruments (Vol. II). Philadelphia: Research for Better Schools, Inc.

Smith, B. O. & Meux, M. O. 1970. A study of the logic of teaching. Chicago: University of Illinois Press.

Soar, R. S. 1968. Optimum teacher-pupil interaction for pupil growth. Educational Leadership Research Supplement, 1, 275-280.

Splinter P. G. 1980. Observation of teaching behavior in physical education with the Physical Education Interaction Analysis System (PEIAS). Experimental research and explorations with respect to reliability and construct validity of categories. University of Amsterdam, The Netherlands. Doctoral dissertation, project No 0255. Splinter, P., Tavecchio, G., Kemper, H., Ras, J., Snel, J. & Verschuur, R. 1979. A Physical Education interaction analysis System - Sensibility of categories to teaching styles. In T. Tammivuori (ed.) Evaluation: International Congress of Physical Education, July 1976. Helsinki: Finnish Society for research in Sports and Physical Education, 64, 69-81.

Stake, R. E. 1967. The countenance of educational evaluation. Teacher College Record, 68 (7), 523-540.

Stake, R. E. 1975a. Evaluating the arts in education: A responsive approach. Columbus, OH: Charles E. Merrill.

Stake, R. E. 1978. The case study method in social inquiry. Educational Researcher, 7, 5-8.

Steward, M. J. 1977. A Descriptive analysis of Teacher behavior and its Relationship to presage and context variables. Doctoral dissertation. The Ohio State University. University microfilms No 48108. Dissertation abstract, No 77 31991.

Stufflebeam, D. L. 1968. Toward a science of educational evaluation. Teachers College Record, 68 (7), 523-540.

Tavecchio, L. W. C. 1977. Quantification of teaching behavior in physical education. University of Amsterdam, Doctoral dissertation, Holland: VRR Groningen.

Tavecchio, L. W. C., Splinter, P. G., Kemper, H. C. G., Koos, G. A., Ras, J. S. & Vershuur, R. 1977. Development and application of physical education interaction analysis system. International Journal of Physical Education, 14 (1), 12-19.

Telama, R. 1967. Jyväskylän yliopiston liikuntakasvatuksen opintosuunnan valintakokeiden valideettitutkimus.

Telama, R. 1968. Liikuntakasvatuksen opiskelijoiden valinnasta. Stadion, (The Finnish Research Association) 5(2), 19-22.

Telama, R. 1975. Education of Physical Education Teachers in Finland and its reform. Reviews of the department of Physical Education, University of Jyväskylä. No 3.

Telama, R. 1978. The meaning of evaluation in the development of Learning and teaching. In M. Piéron (ed.) Towards a Science of teaching Physical Education: Teaching analysis. AIESEP yearbook v.2. Liège: University of Liège: 31-43.

Telama, R. 1970. Teacher's attitudes, expectations and learning results in further training III. Reports from the institute for educational research 65/1970. University of Jyväskylä.

Telama, R. 1979. The training of physical education teachers in Finland. International Journal of Physical Education, 1979; 4, 8-15.

Telama, R. 1990. Problems of meaningfull learning in Physical Education and Paritcularely in Teacher Training. In R. Telama, L. Laakso, M. Piéron, I. Ruoppila and V. Vihko (eds.) Physical Ecucation and Life long learning. Jyväskylä. Reports of Physical Culture and Health No 73, 85-91.

Telama, R., Pirttimäki, R. & Vuolle, P. 1980. The role of audiovisual aids in an institute for physical education teacher training. In G. Schilling & W. Bauer (eds.) Audiovisual Means in Sport. Basel: Birkhaus Verlag, 180-193.

Telama, R., Rantakari, J. & Rauhala, J. 1988. Development of students' values and motives for professional career during 5-years teacher training. Paper presented at AIESEP Congress 26.-31. July 1988.

Telama, R. & Vuolle, P. 1976. Reform of the training of physical education teachers in Finland: Theory and Practice. Paper presented at the International Conference of A.I.E.S.E.P. Bizerte, Tunisia 29.9.-2.10.1976. In T. Haajanen & M. Veistola (eds.) Research in Physical Culture in Finland, Policy in Physical Culture Research work. Abstracts, IV 1976. Reports of the Finnish Society for Research in Sports and Physical Education No 55. Helsinki: Finnish Society for Research in Sports and Physical Education, 56-57.

Tousignant, M. & Brunnelle, J. 1982a. What we have learned from students and how we can use it to improve curriculum and teaching. In M. Piéron & J. Cheffers (eds.) Studing the teaching in Physical education. Liège: A.I.E.S.E.P., 3-22.

Underwood, G. L. 1979. The use of interaction analysis videotape recording in studying teaching behavior in physical education. In T. Tammivuori (ed.) Evaluation: International Congress of Physical Education Jyväskylä, July 1976. Helsinki: The Finnish Society for Research in Sport and Physical Education, 64, 59-69.

Underwood, G. L. 1980. A Comparison of direct and problemsolving approacher in the teaching of physical education. In G. Schillings & W. Bauer (eds.) Audiovisual Means in Sport. Moyen audiovisuelles dans le Sport. Audiovisual means in Sports. Basel: Birkhausen Verlag, 285-296.

Varstala, V. 1990. Teacher behaviour in school physical education classes. In R. Telama, L. Laakso, M. Piéron, I. Ruoppila & V. Vihko (eds.) 1990. Physical Education and Life-Long Physical activity. Report of Physical Culture and Health, 73. Jyväskylä: Foundation for Promotion of Physical Culture and Health, 414-422.

Varstala, V. 1996. Teacher behavior ans students' motor engagement time in school physical education classes. Doctoral dissertation University of Jyväskylä.

Varstala, V., Telema, R. & Akkanen, O. 1981. Teacher and students activities during physical lessons. In Haag et al. (eds.) Physical Education Evaluation. Schorndorf: Hoffmann. Schiffenreihe des Bundesinstitut für Sportwissenschaft, Band 36, 368 – 374.

Varstala, V., Paukku, P. & Telama, R. 1983. Teacher and pupil behavior in physical education classess. In R. Telama (ed.) Research in School Physical Education. Reports of Culture and Health, 38. Jyväskylä, 47 – 57.

Vuoden 1973 opettajankoulutustoimikunnan mietintö (Report of the 1973 Commission on Teacher Education). Komiteanmietintö 1975: 75. Helsinki: Government Printing Office.

Wagner, A. C. 1971. Changing teaching behavior: A comparison of microteaching and cognitive discrimination practice. University of Michigan. Doctoral dissertation. Dissertation Abstract, No DAI. V. 32 A - no 3-1389.

Westbury, I. & Bellack, A. (eds.) 1971. Research into classroom processes: Recent developments and next steps. New York: Teachers College Press.

Whitehead, N. J. 1976. Effecting Change in the Physical Education curriculum. In J.E. Kane (ed.) Curriculum Development in Physical Education. London: Crosby Lockwood Staples, 165-186.

Whitehead, N. J. & Hendry, L. B. 1976. Teaching Physical Education in England. Description and analysis. London: Lepvs Books.

Whitehead, N. J. 1980. The selection of candidates for physical education and sports degree courses. International Journal of Physical Education 17 (4), Winter 1980, 21-23.

Withall, J. 1949. The development of a technique for the measurement of social-emotional climate in classroom. Journal of Experimental Education, 17 (7), 347-361.

Wittrock, M. C. (ed.) 1986. Handbook of research on teaching (Third Edition). New York: Macmillan Publishing Company.

Worthen, B. R. & Sanders, J. R. 1987. Educational Evaluation. Alternative Approaches and Pratical guidelines. New York: Longman, 127-143, 10.

Yerg, B. & Twardy, B. 1982. Relationship of specified instructional teacher behaviors to pupil gain on a motor skill task. In M. Piéron & J. Cheffers (eds.) Studing the teaching in Physical Education. Liège: AIESEP, 61-68.

18 LIST OF FIGURES

19 LIST OF TABLES

20 APPENDICES

APPENDIX 1.2 Procedure of observation

The procedure of observation (PEIAC/LH-75)

The observer places himself where he can hear and see both the teacher and the pupils, or the video recording on the TV monitor. He observes the first five minutes from the beginning of the lesson without marking the card. The observervation period is started and terminated by marking 1287 in the first and last row of the appropriate column. Then every six seconds, either on hearing the signal of by following the hands of the large clock placed on top of the TV receiver, the observer decides which of the three clusters in the classification system the events of the previous six seconds best belong to. The observer writed down the nubers selected while following the events of the next period. Thus he continues for twenty minutes making four digit codings in the appropriate row of the answer card in the six second columns, ten codings per minute. The chronology of the events is retained. A louder signal marks the end of a five minute period, whereupon the observer must continue marking in the first column of the row reserved for the next five minutes.

Where certain events in the observation period have been unclear, this is indicated in the rows (2 vertical lines) at the beginning or end of the said period and a more precise explanation is given at the right-hand edge of the card or on the back. Other features which are necessary for the later interpretation of results are indicated, for example, whether the class was divided, the size of the group observed that was moving etc.

L. Heinilä 1976

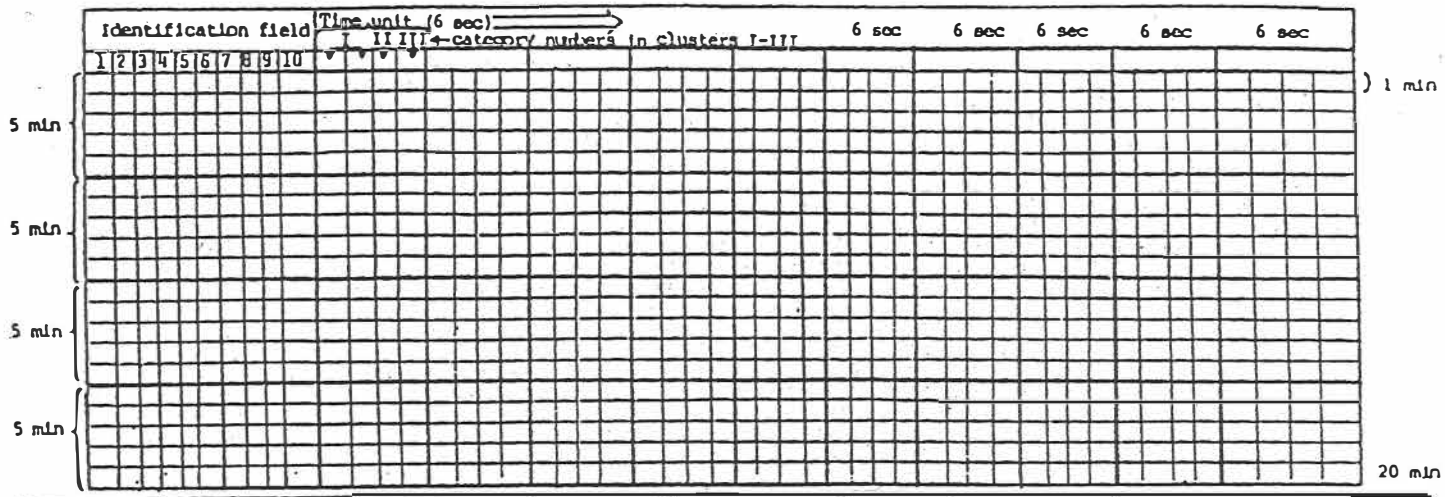
APPENDIX 1.2.1 Classification time sheet

The classification time sheet (see appendix) is the same as an ADP coding sheet where information on the variables connected with lesson material is located in columns 1 – 8, the sequence number of the card in columns 9 – 10, and the observations on the teaching process within the time units in columns 11- 78.

Before the commencement of the observation period the observer fills in information on the factors below in the first ten columns of the time sheet.

Column:		
	1	Observer number (1 – 6)
	2 & 3	Situation 01 – 24
	4	Classification time: 1. natural situation, 2. video-tape, 3. video-tape, 4. sound tape
	5	Measure 1- 9
	6	Class level: 1. preschool, 2. junior comprehensive, 3. intermediate comprehensive, 4. senior comprehensive, 5. sixth form comprehensive, 6. other
	7	Teacher: 1. man, 2. woman
	8	Subject matter: 1. free gymnastics, 2. apparatus, 3. rhythmic movement expression, 4. ball games, 5. basic sport
	9 & 10	Sequence nuber of card
	11 – 80	Variables

APPENDIX 1.2.2 The coding sheet employed in recording

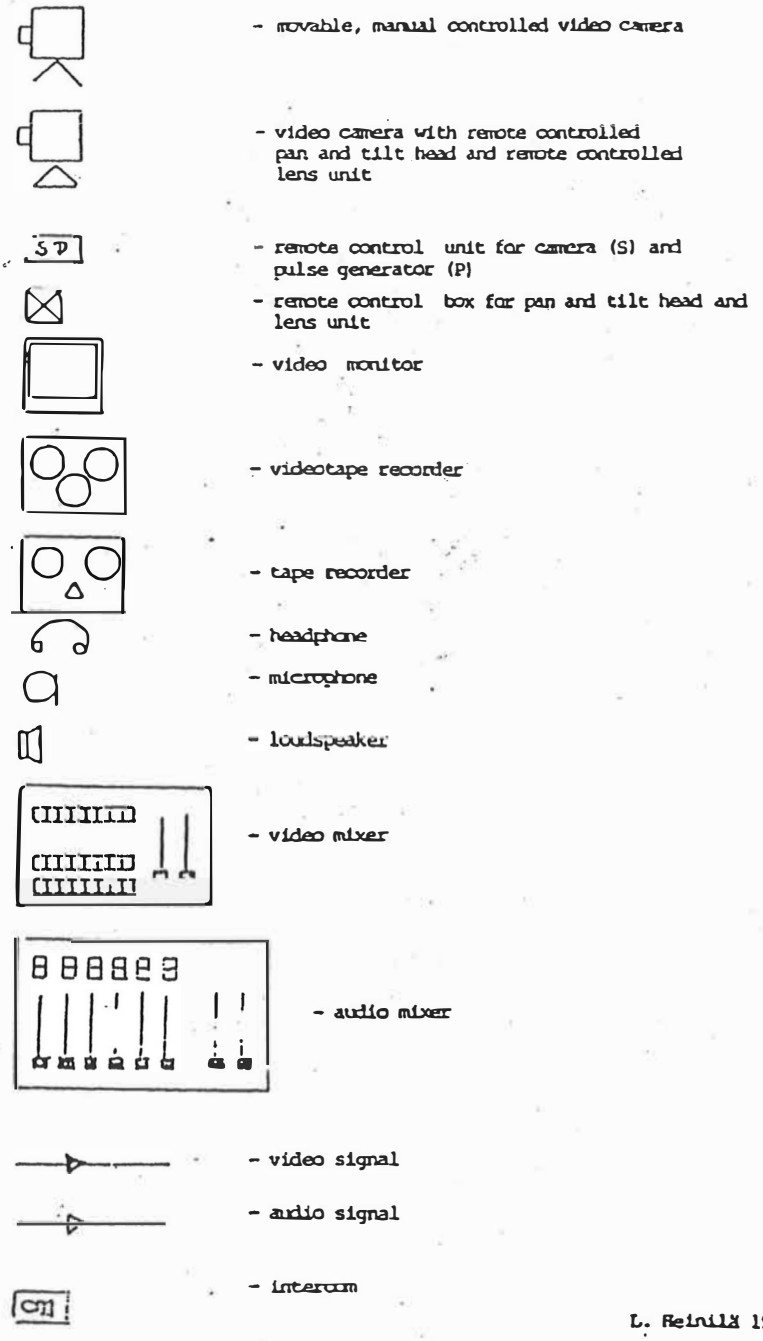


1. Column	2-3. Columns	4. Column	5. Column	6. Column	7. Column	8. Column	9-10. Columns
Coder	Coding situation	Coding occasion	Instrument	Grade level	Teacher	Subject area	Sequence number of punch-card
1 (A) 2 (B) 3 (C) 4 (D) 5 (E) 6 (F)	01 - 24	1 Live situation 2 Video-tape I 3 Video-tape II 4 Sound tape	1 LII/3 clusters 6 sec 2 LII/1.cluster 6 sec 3 LII/2.cluster 6 sec 4 LII/1 & 2 6 sec 5 LII/3 & 2 6 sec 6 LII/2 & 1 6 sec 7 Other	1 Preschool 2 Junior comprehensive 3 Intermediate comprehensive 4 Senior comprehensive 5 Sixth form comprehensive 6 Other	1 Man 2 Woman	1 Gymnastics 2 Apparatus 3 Rhytmic movement expression 4 Ball games 5 Basic sports	01 - 99

L.Heinilä 1976

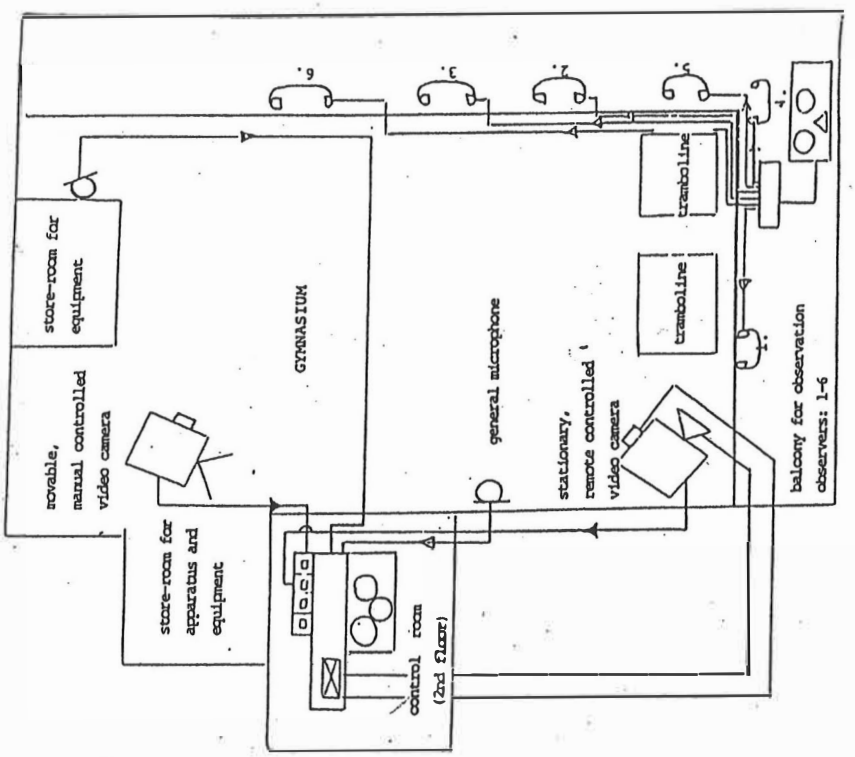
APPENDIX 2.1 Audio-visual equipment and arrangement (1974 -)

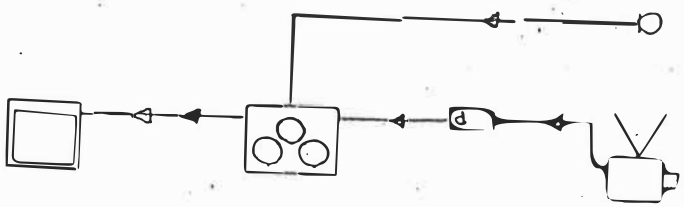
1. Symbols used for technical equipment



APPENDIX 2.1.1 Placement of video cameras, microphones and observers in the gymnasium

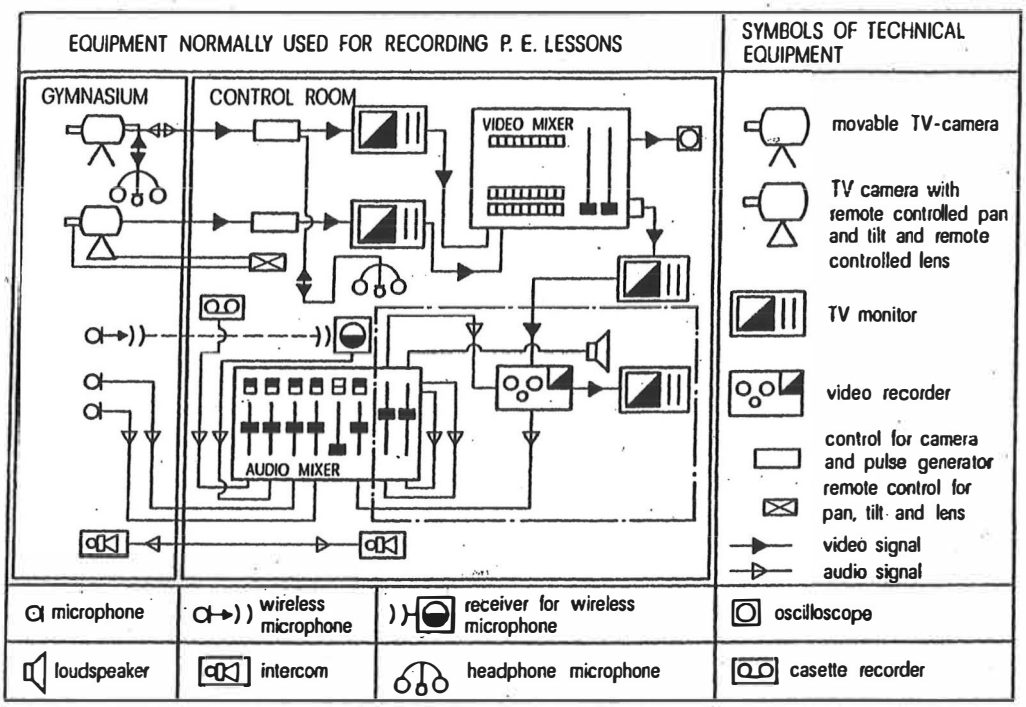
APPENDIX 2.1.2 Scheme of SHIBADEN video equipment (recording)





L. Heinlä 1976

APPENDIX 2.2 ITV (Intern Television System 1980-) in the faculty of health and physical education of the University of Jyväskylä



Telama, R., Pirttimäki, R. & Vuolle, P. 1980

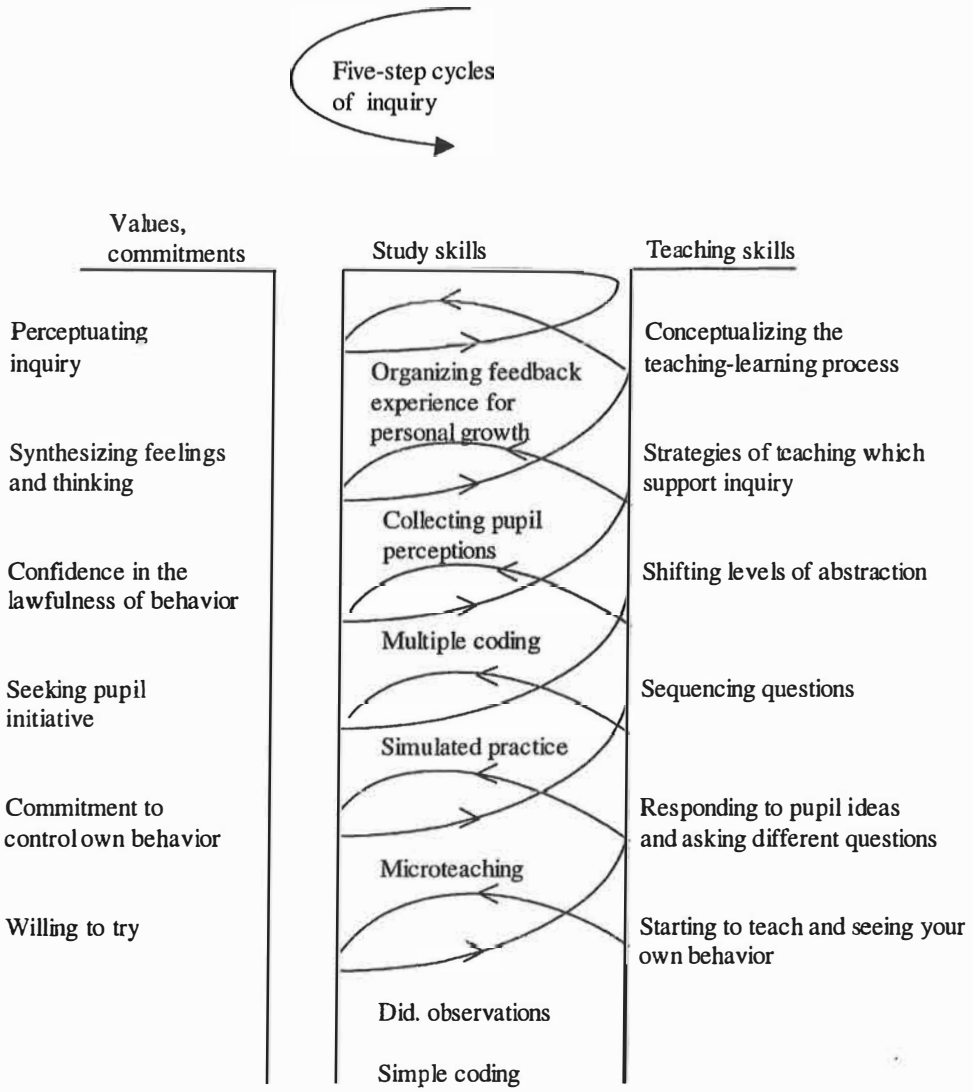
APPENDIX 3 Means and standard deviations of six observer scores by using the PEIAC/LH-75 category system in video-recorded material observation (T_2)

Cluster		Categories	A N=24		B N=24		C N=24		D N=24		E N=24		F N=24		Total N=144	
			X	S	X	S	X	S	X	S	X	S	X	S	X	S
		Teacher's talk, movement, pupils' talk, other														
I	Teacher	01. Accepts, praises	5.46	4.06	4.92	3.53	5.17	4.41	5.63	4.36	5.42	3.75	10.21	5.76	6.13	4.68
		02. Gives corr. feedback	6.33	6.32	12.54	5.80	10.67	10.10	16.58	13.48	11.08	7.73	10.04	8.33	11.21	9.37
		03. Uses ideas dev. by pup.	.17	.48	.88	1.15	.21	.72	.25	.44	1.00	1.78	1.25	1.98	.63	1.19
		04. Asks; init., term. act.	11.00	9.16	12.67	10.45	10.79	8.50	11.83	7.43	13.42	9.84	21.04	13.77	13.56	10.49
		05. Presents inform., org.	79.13	19.37	72.67	19.07	83.92	21.72	79.45	18.03	88.13	22.46	72.25	17.07	79.26	20.19
		06. Gives dir., comm.	8.67	9.21	4.71	7.91	6.58	10.91	7.67	10.09	6.54	8.41	11.95	9.94	7.69	9.48
		07. Criticizes	1.54	2.82	.54	1.28	1.96	2.49	2.38	3.05	1.04	1.52	1.87	2.91	1.56	2.48
	Pupil	08. Answers questions	1.67	1.93	1.08	2.10	.54	1.28	.88	1.36	.92	1.44	2.45	2.67	1.17	1.93
		09. Speaks spontan., init.	1.92	2.70	1.67	2.18	3.58	3.64	3.04	3.54	1.96	2.29	7.79	5.82	3.33	4.11
	Teacher	10. Silent guidance	71.75	25.51	73.21	26.33	63.00	27.00	58.00	29.55	57.12	28.24	48.08	25.14	61.68	27.93
		11. Silent participation	10.71	14.59	13.08	18.30	11.58	18.07	12.25	18.77	11.38	18.34	10.75	17.03	11.63	17.28
	Other	12. Confused situation	2.17	.87	2.04	.20	2.00	0.00	2.04	.46	2.00	0.00	2.29	.91	2.09	.55
II		Pupil's collective movement activity/ passivity and social access														
	Activity	1. Contacts, ideas cont.	22.88	38.99	22.21	34.52	19.79	31.03	20.46	30.80	23.58	35.34	27.83	33.56	22.79	33.65
		2. Contacts free, ideas cont.	85.00	50.03	80.96	47.78	87.08	47.10	81.54	49.69	79.83	51.49	73.58	45.64	81.33	48.00
		3. Contacts free, ideas open	16.63	33.86	16.12	33.01	13.79	30.90	18.46	34.45	16.29	33.54	15.75	31.53	16.17	32.35
		4. Pupils' spont. activity	1.67	3.67	.71	3.07	1.33	3.90	.63	2.86	.71	3.26	1.21	2.99	.96	3.27
	Passivity	5. Pupils follow instruction	49.63	19.97	56.25	22.89	50.96	22.02	56.25	24.08	56.96	23.97	56.29	23.79	54.39	22.62
		6. Pupils organization	21.83	12.16	20.46	10.16	22.75	11.17	19.71	10.98	19.96	11.47	20.92	10.86	20.94	11.10
		7. Pupils wait for turn	9.92	1.95	1.25	1.94	2.29	2.63	.33	.96	.67	1.66	2.25	2.34	1.28	2.09
	Other	8. Confused situation	1.96	.20	2.04	.20	2.00	0.00	2.63	1.24	2.00	0.00	2.17	.56	2.13	.61
III		Social form														
	Situation	1. Complete class, uniform task	59.54	57.24	66.04	54.68	62.08	56.14	63.12	55.42	62.63	56.61	64.25	56.09	62.94	55.08
		2. Divided class, uniform task	56.79	69.99	55.45	70.66	57.00	70.60	50.96	64.61	58.04	70.37	57.79	67.75	56.01	67.86
		3. Divided class, different tasks	46.12	57.08	43.46	54.48	45.96	57.43	47.14	56.72	45.25	56.13	43.54	57.90	45.25	55.65
		4. Div. cl. diff. task within gr.	20.50	34.94	16.58	29.98	16.85	30.71	20.04	32.55	17.29	31.32	16.45	29.88	17.96	31.10
		5. Indivudual work, unif. tasks	13.83	27.78	15.33	30.86	14.83	29.85	15.58	28.72	14.08	28.34	15.00	30.46	14.78	28.85
		6. Individual work, diff. tasks	.58	2.86	.46	2.24	.54	2.65	1.04	5.10	.04	.02	.37	1.84	.51	2.84
		7. Other, conf. situation	2.63	3.06	2.67	3.06	2.71	3.48	2.08	.41	2.67	32.67	2.58	2.86	2.56	2.84

6 observers (A-F)
24 lessons
4800 6 second time units, tot. 28800 time units

APPENDIX 4 **The main elements of the study unit course of didactic observation and microteaching 1976 –**

APPENDIX 4.1 Curriculum elemts in professional self-development (after Hilda Taba) (Flanders 1970, 306, Heinilä 1977b)



APPENDIX 4.2 University of Jyväskylä/Department of Physical Education, information:

APPENDIX 4.2.1 Course of didactic observation

University of Jyväskylä/Department of Physical Education

Course of didactic observation (I) and lectures and practice microteaching (II):
1974-1979 /cl 4c, cl 4d(30 h; 15 h + 15 h; 45 h, 15 + 30 h)
1980-1991 (two study weeks) (15 h + 12 h; 15 h + 20 h)

I COURSE ON DIDACTIC OBSERVATIONS

Objectives:

Lectures

To provide students with:

1) knowledge of general theory and basic concepts of education, and different elements in teaching-learning process; as well as ability to understand, analyse and evaluate this knowledge,
2) ability to systematically observe and evaluate one's own, and also others' teaching behavior,
3) ability to construct different frames of references using coding systems; as well as ability to analyse, present and interpret results of systematic observation,
4) willingness to develop and widen one's personal teaching behavior

Demonstrations

To provide students with:

1) knowledge and technical mastery of systematic observation methods,
2) readiness in objective comparison and evaluation of teaching-learning processes with reference to research,
3) skills how to make use of and present information obtained by systematic observation

Contents

Lectures

1. Starting point, background and definition of tasks
2. Observation research
3. Observation of teacher-pupil interaction in physical education and construction of a method of analysis
4. Observation system: PEIAC/LH-75
   - 4.1 Frame of reference,
   - 4.2 Classification system for physical education interaction process
   - 4.3 Definition of clusters and instruction for classification
   - 4.4 The procedure of observation

Demonstrations and practice

1. Practices in the procedure of observations
   - 1.1 Systematic description of various situations using audiotape, videotape, natural situation
   - 1.2 Coding practices using different coding systems
     - 1.2.1 Observations and coding of movement behavior from manuscript, audiotape, videotape, natural situation
     - 1.2.2 Coding of speech behavior
     - 1.2.3 Practices in microteaching
   - 1.3 Practices in analysing the data
     - 1.3.1 Tabulation

1.3.2 Counting of parameters (primary information)
1.3.3 Presentation of results (profile, time-line display, matrix)
1.3.4 Counting of reliability
1.4 Interpretation and conclusions
1.5 How to construct coding systems
1.6 Practice study (comparison to two teaching situations)
1.7 Evaluation of practice studies experi-mentations (reports)

Form of the course

Timing: before study reform (1978) organized organization in the term of the third year and after study reform in the least term of the second study year.
Obligatory course for all students grouped 15-20 persons.

Material

1. Hand-outs (scetch for lecture, different coding systems)
2. Methods for reliability determination
3. Research reports
3. Transparents for over-head projector; all the main points of the lecture are presented in transparents including paradigms, models, frames of references, research results
3. Audio- and videotapes of teaching processes.
During the lectures, especially in the beginning, the problems to be studied and evaluated are visualized with the help of audio- and videorecordings, in short 2-5 min periods.

Evaluation:

Cl, 4c, theory-test (1-3 p), practice (1-3 p), total scores 1-3 p. After studiereform one study week (1-3 p) combined to the evaluation of microteaching course (50%, 50%) 1-3 p.

Lectures:

1. 50% of time used on lecturing consists of theory and background of the observation system used, its application to teacher training, basic concepts and elements used in microteaching course and program evaluation.
2. 10% of time is spent in model demonstrations
3. 25% of time in is practice from which 50% is systematic observation, coding, analyzing and evaluation of teaching process.
4. 15% of time is spend in group work for practice study and its presentation and evaluation.

APPENDIX 4.2.2 Course of microteaching

II COURSE ON MICROTEACHING

The main elements of the course can be briefly described as follows:

Objectives

Lectures:
To provide student with: knowledge and mastery as well as cognitive understanding of characteristics of indirect verbal and non-verbal behavior in P.E. as defined in author's adaptation of Flanders' Interaction Analysis System PEIAC/LH-75 I (Heinilä 1977b).

Contents	1. Theoretical background Microteaching and of selected teaching models (1-6) 2. Instrument of observation PEIAC/LH-75 II and its use. Reliability control. 3. Model demonstrations (1-6) in different settings. 4. Evaluation of teaching event, discrimination of model. 5. Students' attitudes, "Ideal" P.E. teacher expectation inquiry analysis. Results. Comparison with earlier results obtained.
Objectives	Practice: To provide students with: 1. mastery and cognitive understanding of on a non-directive teaching behavior, by exercising teaching models, 2. readness in planning of teaching events in pre-interactive phase of microteaching in frames of teaching strategy model, 3. readness in reflective flexible teaching behavior in microlessons with frame of teaching strategy model, 4. readness in observation analyzing and evaluation of teaching events based on frame of reference, 5. willingness to develop and widen one's personal teaching behavior and use of non-directive teaching: readness in reflective teaching.
Contents	Practice: information, teach one 5 min (control) evaluation videotape replay by using scale 4 items (1-6), planning of microlesson one, teach one (10 min) videotape replay, self-observation (10 min) by using PEIAC/LH-75 II, analysis, evaluation and discussion (10 min); replanning, microlesson two, reteaching (10 min), videotape replay (10 min) self-observation, analysis, evaluation and comparison of microlessons one and two; summative evaluation (58 items, 1-5 scale)
Form of the course and organization	Timing: before studyreform (1978) in the second term of the third study year; after studiereform in the first term of third year. Group membership: students could select own group-membership (7-10 students) also in mixed gender-group. The other members of the course group (n=7-9) will served as pupils during microlessons, observed the lessons given by all other students from video tape replay, and taked part in the analysis and discussions, and one student is the principle opponent.
Material	1. Hand-outs (catch for lecture and practice) 2. Task plan teaching strategy model: timing, framefactors (teaching model, 3. Subject are, pupil age level readness, cabability, lesson plan form. 4. Observation instrument (PEIAC/LH-75 II) and information of its use, coding sheet, time-line display. 5. Model demonstration videotapes. 6. Questionnaire for program evaluation. 7. Questionnaire for students "Ideal" P.E. teacher expectations (20 items, 1-6) raport.
Evaluation	Theory (1-3), practice (1-3), total scores (1-3), after studiereform evaluation of total scores was combined with the total scores of the course of didactic observation (50%, 50%, two study week, 1-3 p).

APPENDIX 4.2.2.1 Instruction for a practice study (report)

Liisa Heinilä 1978
University of Jyväskylä
Finland

INSTRUCTIONS FOR A PRACTICE STUDY (REPORT)
WHERE SYSTEMATIC OBSERVATION METHOD IS USED

Practice study is a proof of the physical education student's ability to analyse, by using systematic observation method of the teaching-learning process. Also the student should learn construct a comparative research strategy, and to interpret and present the results in a form of a report.

In the practice study you describe and compare two teaching situations from the point of view of the problem, and the frame of reference chosen as to your preference. For collection of the study data a systematic observation method is used. The aim of your study may also be to evaluate the observation instrument you have used. For example, you compare the applicability of the observation instrument in two situations: 1) observation from the videotape and 2) observation from the "natural situation" (direct observation). You are asked to present a written report of your practice study at the end of the course.

The practice study is made in groups composed of two persons. However, each member of the group observes individually. The study group makes decisions, together, for the following stages of planning and realization: 1) the choice of the problem, 2) the frame of reference, 3) the variables, i.e. the things to be observed and the category system, 4) the choice of the sample unit (i.e. time interval of coding), 5) the procedure of observation, 6) the data (situations), 7) analysing the data, 8) the analysing methods, 9) the interpretation and presentation of the results.

When choosing the setting for comparison (i.e. situations, categories, variables) it is advisable to make sure that only one independent variable is chosen, e.g. teacher, subject matter, sex of the pupils, grade level, a certain period of the class or form of teaching. The samples should be two twenty minute periods. The same teaching situation may be re-observed, e.g. in natural situation, from vidcotape or from soundtape.

Reliability of coding is studied by assessing the agreement of the entries (markings in the categories) of the two observers who have coded the situation simultaneously.

The chosen analysing- and categorysystem may be any one introduced during the lecture or an applied, expanded or shortened form of the sytem. The source and origin of the system must be mentioned and the reason for choice explained. You may as well construct a system of your own.

When constructing a coding system attention must be paid to the following poits:

1) What is included in the problem you are going to observe.
2) What are the concepts necessary for analysing and explaining the typical features of this problem.
3) How can you operationalize these concepts (i.e. change them into category definitions).
4) How can you measure and quantify them by using a systematic observation method (choice of time interval and observation period).
5) Which theory or research could best throw light into the relationships of these concepts.

Teacher talk	Response	1.	Praises, encourages, accepts the feeling tone of a pupil
		2.	Gives corrective feedback, directs, clarifies, answers pupil's questions
		3.1.	Makes use of the ideas and movement patterns suggested by a pupil: clarifies, expands, builds questions and movement initiations on the ideas expressed by a pupil
		3.2.	Summarizes pupil's ideas or movement patterns, asks a pupil to demonstrate
		3.3.	Compares the ideas or movement patterns expressed by one pupil to those of another or to those given, repeats pupil's ideas, asks a pupil to demonstrate
	Question	4.1.	Asks questions requiring narrow answers, initiates short-term activity, terminates activity
		4.2.	Makes questions requiring higher level of thinking or activity
	Initiation	5.1.	Presents information, opinions, demonstrates movement patterns, makes a pupil demonstrate
		5.2.	Organizes pupils, material, division of labour and responsibility
		6.	Gives directions, commands during activity (pupils expected to comply)
		7.	Criticizes pupil behaviour, rejects movement pattern, justifies authority
Pupil talk	Resp.	8.	Pupil answers question made by the teacher
	Init.	9.	Pupil initiates speech, asks for instructions, expresses own ideas or movements
Other	Silence, confused situation	10-12	(10) Teacher follows pupils' activity, silent guidance (11) Teacher's silent participation in movement activity (12) Confused situation, uproar
The decision on classification is made on the basis of the didactic function of the activity.			

APPENDIX 4.2.2.2.1 Coding instructions and coding sheet

Instructions for Classification

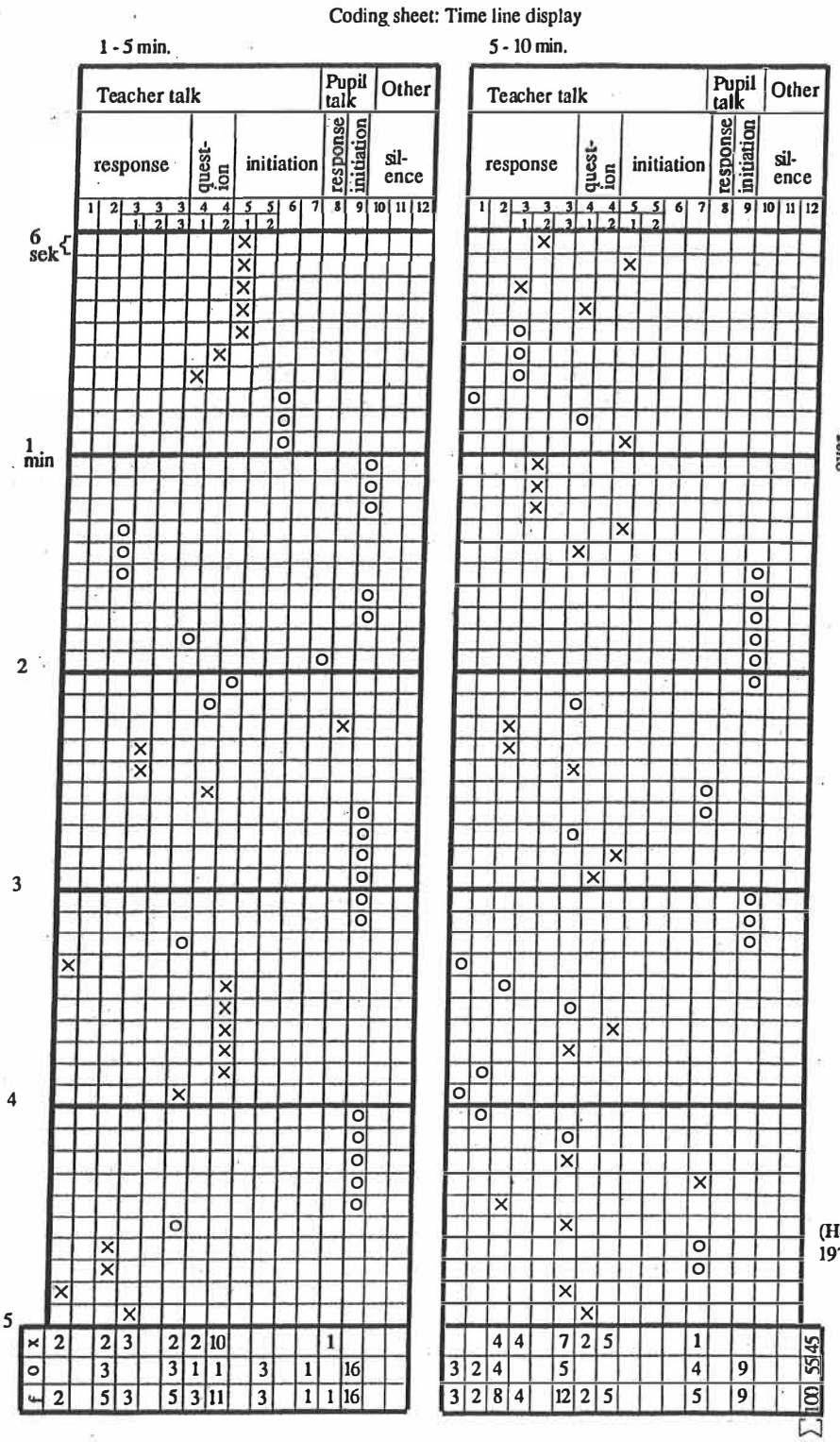
Before the beginning of the observation perid the observer enters on the reverse side of the form data on teaching situation.

The observer places himself where he can hear and see well the TV-display. Every sixth second, either on hearing a signal or observing the clock placed on topof the TV set, he decides which of the categories of the classification system best represents the events of the previous six-second period. The observer directs his attention to the speech and movement behaviors of the teacher and students. Students' movement behavior is viewed collectively. The observer marks the relevant category column entering either O or X depending on whether the class was active or passive in terms of movement during the six-second period. At the same time he observers what is happening during the next period. This produces 10 entries per minute and 100 entries in ten minutes. At every full minute timing should be checked. Entries on the form constitute a series going from top to bottom, which preserves the sequence of events. Categories have been placed on the form so that entries yield an immediate basis, for example, for (I) a visual evaluation of teacher's initiating and response behavior, (II) a general idea of the amount of movement, (III) a general idea of the stability and variability of the process, and (IV) a general picture of nature of points.

At the end of the observation period, column totals are computed and entered on the form, separately of O and X and combined. The column totals of O, X and F are also added up. After that computations for obtaining indices (on the back page of the form) are carried out. The obtained results are used in analysing, comparing and evaluating micro lessons in relation to set objectives.

(Heinilä, 1977)

APPENDIX 4.2.2.2.2 Coding sheet: time line display



APPENDIX 4.2.3 Microteaching course/Heinilä, L. 1975 (feedback, intervention sheet for lessons)

Microteaching course/Heinilä, L. -75

Class n:o 1 2

Student (subject) n:o ___ Name: _____ Model n:o _____ Date: _____

Class information: age of pupils ___y. skill level ___ Subject matter _____

Calculate the following indices:

1) Percent teacher talk $= \frac{100 - (8+9+10+11+12)}{100} \times 100$

2) Percent pupil activity $= \frac{100 - \Sigma X}{100} \times 100$

3) Teacher response ratio $= \frac{1+2+3+11}{1+2+3+6+7+11} \times 100$

4) Occurance of models (frequencies)

5) Percent model occurance $= \frac{\text{categories in the model}}{100} \times 100$

6) Intensity of teacher guidance $= \frac{4+6}{1+2+3+4+5+6+7} \times 100$

Observation instructions for teaching models:

Way of teaching	Model n:o	Category n:o
Direct teaching		
Information presentation model	1.1.	5.1., 4, 6
Organization model	1.2.	5.2., 4, 6
Initiation variation model	1.3.	4.1., 4.2., 5.1.
Indirect teaching		
Teacher initiations based on pupils responses	2.1.	3.1.
Reinforcing pupil initiations, summarizing model	2.2.	3.2.
Comparison making model	2.3.	3.3.
Accepting pupil's feeling -model	2.4.	1, 2
Corrective feedback -model	2.5.	2, 3.1.
Changing the level of abstraction	2.6.	1, 2, 3, 5.1.
Teacher's reinforcing and extinguishing reactive behavior	2.7.	1, 2, 3, 5.2.
Discrimination making model	2.8.	2, 3, 4, 5

Analyse: _____

_____

Suggestions for improvements: _____

_____

_____ _____

date signature

APPENDIX 5 Student program evaluation (1)

APPENDIX 5.1 Student evaluation of instruction, questionnaire

STUDENT EVALUATION OF INSTRUCTION

With this questionnaire you have an opportunity of stating your views about the instruction you have received. Your opinions are valuable, since they can be used for improving the course. Please, answer all questions. You can answer anonymously.

Answer by circling for each question or statement the number of the alternative that best fits your view.

1. Sex Male = 1 Female = 2 (1)

Below you can read a number of statements. Answer each of them in the following way:
if you disagree completely with a statement, circle number 1
if you disagree to some extent, circle number 2
if you are uncertain of feel it does not matter, circle number 3
if you agree to some extent, circle number 4
if you agree completely, circle number 5

2. The course was pretended so that I was aware of its contents and extent from the beginning.....	1	2	3	4	5	(2)
3. I was able to get the right idea of the objectives of the lecture course from the beginning.....	1	2	3	4	5	(3)
4. I was aware of the objectives of the exercises from the beginning.....	1	2	3	4	5	(4)
5. The main concepts of the course were badly presented.....	1	2	3	4	5	(5)
6. The course has awoken an interest in me in this subject.....	1	2	3	4	5	(6)
7. I did not learn much during the lectures	1	2	3	4	5	(7)
8. I did not learn much during the exercises	1	2	3	4	5	(8)
9. Exercise tasks have been sensible.....	1	2	3	4	5	(9)
10. Using students as "pupils" has been reasonable	1	2	3	4	5	(10)
11. The course has overlapped unnecessarily with my earlier studies.....	1	2	3	4	5	(11)
12. The course was sensible linked with earlier studies.....	1	2	3	4	5	(12)
13. The course was organized well compared with other corresponding courses.....	1	2	3	4	5	(13)
14. The contents of the lecture and the exercises did not match each other sufficiently.....	1	2	3	4	5	(14)
15. This course should have been placed earlier in the study programme.....	1	2	3	4	5	(15)
16. Exercises contained too few tasks of different types.....	1	2	3	4	5	(16)
17. Exercises proceeded too quickly.....	1	2	3	4	5	(17)
18. Too little time was spent on the analysis of feedback.....	1	2	3	4	5	(18)
19. Lectures should have included more audiovisual equipment.....	1	2	3	4	5	(19)
20. Lecturer has been to detached (impersonal).....	1	2	3	4	5	(20)
21. Lecturer has spoken loud enough...	1	2	3	4	5	(21)
22. Throughout the semester I remained unware of the objectives of the course	1	2	3	4	5	(22)
23. The main concepts of the course have been presented clearly enough.....	1	2	3	4	5	(23)
24. Lecturer did not give the students enough time to ask questions.....	1	2	3	4	5	(24)
25. During the course teachers were careless as regards deadlines for assignments.....	1	2	3	4	5	(25)
26. I was generally bored during lectures	1	2	3	4	5	(26)
27. I was generally bored during exercises	1	2	3	4	5	(27)
28. Handouts outlining the contents of lectures were useful from the point of view of attaining the objectives of lectures.....	1	2	3	4	5	(28)
29. It was difficult to follow the lecture	1	2	3	4	5	(29)
30. The whole course is useless in educating P.E. teachers.....	1	2	3	4	5	(30)
31. Lecturer should have proceeded more quickly.....	1	2	3	4	5	(31)
32. Lecturer did not know the subjects well enough.....	1	2	3	4	5	(32)

APPENDIX 5.1 continued

33. Lecture's personal opinions biased teaching too much.....	1	2	3	4	5	(33)
34. Time reserved for exercises was usually too short.....	1	2	3	4	5	(34)
35. The course as such is rather useful...	1	2	3	4	5	(35)
36. The course did not deal with really essential and important matters.....	1	2	3	4	5	(36)
37. Lectures and exercises were integrated well.....	1	2	3	4	5	(37)
38. It was easy to keep interested in the subject during the lectures.....	1	2	3	4	5	(38)
39. It was easy to keep interested in the subjects during the exercises.....	1	2	3	4	5	(39)
40. The course did not awake any interest in me in the subject.....	1	2	3	4	5	(40)
41. Lecturer has pointless habits and mannerisms which divert the students' attention from teaching.....	1	2	3	4	5	(41)
42. This course should have been placed later in the study programme.....	1	2	3	4	5	(42)
43. I have learnt more in the lectures of this course than in lectures in general	1	2	3	4	5	(43)
44. Lecturer did not take the students into consideration well enough.....	1	2	3	4	5	(44)
45. Handouts summarizing the main points of lectures were useless.....	1	2	3	4	5	(45)
46. Lecture course gave me new ideas about P.E. teaching.....	1	2	3	4	5	(46)
47. Demonstration tasks were badly selected	1	2	3	4	5	(47)
48. Lecture course was not worth attending	1	2	3	4	5	(48)
49. From the point of view of educating P.E. teachers it would have been more useful to spend the time on other types of teaching practise.....	1	2	3	4	5	(49)
50. Demonstrations of lecture and teaching models would be sufficient without having to participate in exercises.....	1	2	3	4	5	(50)
51. Lecturer proceeded too quickly.....	1	2	3	4	5	(51)
52. Organization of exercises was not good enough.....	1	2	3	4	5	(52)
53. The teaching skills of the exercise supervisor were not good enough.....	1	2	3	4	5	(53)
54. The exercise tasks were explained clearly.....	1	2	3	4	5	(54)
55. I learned to distinguishes teaching models observing and classifying feedback....	1	2	3	4	5	(55)
56. Exercises clarified the issues presented in lectures.....	1	2	3	4	5	(56)
57. I believe I have obtained a broader view of teaching behaviour.....	1	2	3	4	5	(57)
58. I will probably use the various teaching models presented consciously in my teaching.....	1	2	3	4	5	(58)
59. I became aware of my personal teaching defects inadequates during the course	1	2	3	4	5	(59)
60. Microteaching should be used also for practising direct models of teaching...	1	2	3	4	5	(60)
61. The task of evaluating teaching was useful.....	1	2	3	4	5	(61)
62. The actual teaching of the planned teaching episode, when only the goal was given, was interesting.....	1	2	3	4	5	(62)
63. Filling in the structural outline for a teaching episode was useless.....	1	2	3	4	5	(63)
64. The way the course groups were set up was sensible.....	1	2	3	4	5	(64)
65. The structural outline facilitated the construction of the plan for the teaching episode.....	1	2	3	4	5	(65)
66. Practice sessions should be carried out in mixed male-female groups.....	1	2	3	4	5	(66)

APPENDIX 5.2 Comparison of curriculum groups 1974 and 1976 evaluation of microteaching course on the percentage items

	1974 N=48			1976 N=73			Total N=121			74-76	
	Dis-agr. %	Un-cer. %	Agr. %	Dis-agr. %	Un-cer. %	Agr. %	Dis-agr. %	Un-cer. %	Agr. %	Diff. df=2 X^2	
2. The course was pretended so that I was aware of its contents and extent from the beginning	52.1	10.4	37.5	60.3	8.2	31.5	57.2	9.1	33.9	.80	
3. I was able to get the right idea of the objectives of the lecture course from the beginning	47.9	22.9	29.2	60.3	12.3	27.4	55.4	16.5	28.1	2.80	
4. I was aware of the objectives of the exercises from the beginning	29.2	8.3	62.5	48.0	4.1	48.0	40.5	5.8	53.7	4.56	
5. The main concepts of the course were badly presented	56.3	33.3	10.4	43.8	17.8	38.4	48.8	24.0	27.3	12.12	**
6. The course has awoken an interest in me in this subject	31.3	18.8	50.0	17.8	15.1	67.1	23.1	16.5	60.3	3.91	
7. I did not learn much during the lectures	37.5	23.0	39.6	48.0	19.2	32.9	43.8	20.7	35.5	1.28	
8. I did not learn much during the exercises	68.8	6.3	25.0	82.2	2.7	15.1	76.9	4.1	19.0	3.05	
9. Exercise tasks have been sensible	18.8	23.0	58.3	41.1	8.2	50.7	32.2	14.1	53.7	9.25	**
10. Using students as "pupils" has been reasonable	70.8	10.4	18.8	39.7	4.1	56.2	52.1	6.6	41.3	16.9	***
11. The course has overlapped unnecessarily with my earlier studies	68.8	22.9	8.3	89.0	8.2	2.7	81.0	14.1	5.0	7.75	*
12. The course was sensible linked with earlier studies	27.1	35.4	37.5	21.9	19.2	58.9	24.0	25.6	50.4	5.94	
13. The course was organized well compared with other corresponding cource	43.8	33.3	23.0	32.9	38.4	28.8	37.2	36.4	26.4	1.50	
14. The contents of the lecture and the exercises did not match with each other sufficiently	45.8	22.9	31.3	61.6	9.6	28.8	55.4	14.9	29.8	4.83	
15. This course should have been placed earlier in the study programme	14.6	20.8	64.6	60.3	13.7	26.0	42.2	16.5	41.3	25.7	***
16. Exercises contained too few tasks of different types	27.1	27.1	45.8	53.4	6.9	39.8	43.0	14.9	42.2	12.9	**
17. Exercises proceeded too quickly	54.2	14.6	31.3	69.9	5.5	24.7	63.6	9.1	27.3	4.22	
18. Too little time was spent on the analysis of feedback	29.2	22.9	47.9	50.7	4.1	45.2	42.2	11.6	46.3	12.08	**
19. Lectures should have included more audiovisual equipment	20.8	49.7	31.3	46.6	16.4	37.0	36.4	28.9	34.7	15.47	***
20. Lecturer has been to detach (impersonal)	50.0	41.7	8.3	78.1	20.6	1.4	67.0	29.0	4.1	11.27	**
21. Lecturer has spoken loud enough	33.3	27.1	39.6	48.0	6.9	45.2	42.2	14.9	43.0	9.65	**
22. Throughout the semester I remained unware of the objectives of the course	68.8	16.7	14.6	68.5	5.5	26.0	68.6	9.9	21.5	5.42	*

TABLE continued

	1974 N=48			1976 N=73			Total N=121			74-76	
	Dis-agr. %	Un-cer. %	Agr. %	Dis-agr. %	Un-cer. %	Agr. %	Dis-agr. %	Un-cer. %	Agr. %	Diff. df=2 X^2	
23. The main concepts of the course have been presented clearly enough	16.7	35.4	47.9	28.8	15.1	56.2	24.0	23.1	52.9	7.32	*
24. Lecturer did not give the students enough time to ask questions	37.5	35.4	27.1	49.3	30.1	20.6	44.6	32.2	23.1	1.69	
25. During the course teachers were careless as regards deadlines for assignments	66.7	14.6	18.8	75.3	9.6	15.1	71.9	11.6	16.5	1.17	
26. I was generally bored during lectures	20.8	33.3	45.8	41.1	15.1	43.8	33.1	22.3	44.6	7.95	*
27. I was generally bored during exercises	79.2	4.2	16.7	74.0	5.5	20.6	76.0	5.0	19.0	.43	
28. Handouts outlining the contents of lectures were useful from the point of view of attaining the objectives of lectures	4.2	4.2	91.7	6.9	1.4	91.8	5.8	2.5	91.7	1.27	
29. It was difficult to follow the lecture	35.4	27.1	37.5	24.7	13.7	61.6	28.9	19.0	52.1	7.13	*
30. The whole course is useless in educating P.E. teachers	83.3	6.3	10.4	79.5	11.0	9.6	81.0	9.1	9.9	.78	
31. Lecturer should have proceeded more quickly	45.8	45.8	8.3	79.5	13.7	6.9	66.1	26.5	7.4	16.34	*
32. Lecturer did not know the subjects well enough	66.7	27.1	6.3	82.2	13.7	4.1	76.0	19.0	5.0	3.92	
33. Lecture's personal opinions biased teaching too much	25.0	54.2	20.8	45.2	23.3	31.5	37.2	35.5	27.3	12.16	*
34. Time reserved for exercises was usually too short	16.7	10.4	72.9	35.6	4.1	60.3	28.1	6.6	65.3	6.15	*
35. The course as such is rather useful	10.4	12.5	77.1	5.5	6.9	87.7	7.4	9.1	83.5	2.36	
36. The course did not deal with really essential and important matters	52.1	27.1	20.8	54.8	11.0	34.3	53.7	17.4	28.9	6.18	
37. Lectures and exercises were integrated well	25.0	29.2	45.8	37.0	13.7	49.3	32.2	19.8	48.0	4.86	
38. It was easy to keep interested in the subjects during the lectures	53.3	29.2	12.5	68.5	16.4	15.1	64.5	21.5	14.1	2.78	
39. It was easy to keep interested in the subjects during the exercises	16.7	4.2	79.2	19.2	5.5	75.3	18.2	5.0	76.9	.26	
40. The course did not awake any interest in me in the subject	54.2	29.2	16.7	64.4	17.8	17.8	60.3	22.3	17.4	2.20	
41. Lecturer has pointless habits and mannerism which divert the student's attention from teaching	27.1	52.1	20.8	43.8	32.9	23.3	37.2	40.5	22.3	4.90	
42. This course should have been placed later in the study programme	77.1	20.8	2.1	75.3	16.4	8.2	76.0	18.2	5.8	2.20	
43. I have learnt more in the lectures of this course than in lectures in general	68.8	31.3	0.0	57.5	31.5	11.0	62.0	31.4	6.6	5.85	

TABLE continued

	1974 N=48			1976 N=73			Total N=121			74-76	
	Dis-agr. %	Un-cer. %	Agr. %	Dis-agr. %	Un-cer. %	Agr. %	Dis-agr. %	Un-cer. %	Agr. %	Diff. df=2 X^2	
44. Lecturer did not take the students into consideration well enough	29.2	31.3	39.6	56.2	17.8	26.0	45.5	23.1	31.4	8.60	*
45. Handouts summarizing the main points of lectures were useless	87.5	6.3	6.3	97.3	1.4	1.4	93.4	3.3	3.3	4.47	
46. Lecture course gave me new ideas about P.E. teaching	16.7	33.3	50.0	20.6	13.7	65.8	19.0	21.5	59.5	6.63	*
47. Demonstration tasks were badly selected	70.8	22.9	6.3	61.6	17.8	20.6	65.3	19.8	14.9	4.74	
48. Lecture course was not worth attending	43.8	35.4	20.8	67.1	20.6	12.3	57.9	26.5	15.7	6.49	*
49. From the point of view of educating P.E. teachers it would have been more useful to spend the time on other types of teaching practise	58.3	16.7	25.0	71.2	9.6	19.2	66.1	12.4	21.5	2.36	
50. Demonstrations of lecture and teaching models would be sufficient without having to participate in exercises	89.6	10.4	0.0	91.8	1.4	6.9	90.9	5.0	4.1	8.08	*
51. Lecturer proceeded too quickly	29.2	43.8	27.1	43.8	26.0	30.1	38.0	33.1	28.9	4.48	
52. Organization of exercises was not good enough	33.3	18.8	47.9	68.5	11.0	20.6	54.6	14.1	31.4	14.72	**
53. The teaching skills of the exercise supervisor were not good enough	47.9	37.5	14.6	83.6	11.0	5.5	69.4	21.5	9.1	17.43	**
54. The exercise tasks were explained clearly	31.9	21.3	46.8	28.8	12.3	58.9	30.0	15.8	54.2	2.31	
55. I learned to distinguishes teaching models observing and classifying feedback	35.4	18.8	45.8	15.1	2.7	82.2	23.1	9.1	67.8	19.00	**
56. Exercises clarified the issues presented in lectures	20.8	29.2	50.0	16.4	15.1	68.5	18.2	20.7	61.2	4.71	
57. I believe I have obtained a broader view of teaching behaviour	8.3	14.6	77.1	5.5	6.9	87.7	6.6	9.9	83.5	2.49	
58. I will probably use the various teaching models presented consciously in my teaching	8.5	12.8	78.7	4.1	13.7	82.2	5.8	13.3	80.8	1.01	
59. I became aware of my personal teaching defects inadequates during the course	19.2	12.8	68.1	12.3	9.6	78.1	15.0	10.8	74.2	1.54	

*, **, ***, p<0.05, 0.01 and 0.001 respectively

APPENDIX 6 Validity of the PEIAC/LH-75 II system

APPENDIX 6.1 Reliability of PEIAC/LH-75 II (2)

Reliability of PEIAC/LH-75 II case study 1988

Objectivity of coding was estimated by means of Scott's coefficient (Pi) between outside trained observer and investigator by using videorecorded microlessons before starting the coding of 1988 case study (n = 42). A sample of 13 cases x 20 min = 260 min was observed these (24.1.1989 and 6.2.1989). The reliability was estimated separately for the two clusters. In the table are summarized the results, Pi values (54) ranged as follows:

	Cluster I (verbal)	MD	Cluster II (movement)	MD
Inter-coder agreement	(.66 - .98)	.76	(.72 - .97)	.84
Within coder constancy	(.79 - .100)	.91	(.92 - .100)	.98
Between coder constancy	(.65 - .85)	.73	(.55 - .98)	.93
	Total MD	.82	Total MD	.91

APPENDIX 6.2.1 Means, standard deviations, and t-tests for the PEIAC/LH-75 II categorie across the three microlessons (n = 221)

	CLUSTER I - TEACHER TALK - PUPIL TALK - SILENT TEACHER ACTIVITY	Categories provided in modified PEIAC/LH-75[1)]	Curriculum 1976 1st Lesson N=73 $\bar{x}$ %	S.D	1st Lesson N=74 $\bar{x}$ %	S.D	2nd Lesson N=74 $\bar{x}$ %	S.D	Total N=221 $\bar{x}$ %	S.D	1 - 2 df=145 t	1 - 3 df=145 t	2 - 3 df=146 t
	No												
Response	01	Praises, encourages, accepts the feeling tone of a pupil	2.1	2.8	5.1	3.9	4.6	3.4	4.0	3.6	5.3***	4.9***	-.75
Response	02	Gives corrective feedback, directs, clarifies, answers pupil's questions	2.8	2.7	5.1	4.7	5.1	4.8	4.4	4.3	3.62***	3.52***	-.02
Response		Makes use of the ideas and movement patterns suggested by a pupil or group of pupils:											
Response	31	Clarifies, expends, builds questions and movement initiations on the ideas expressed by a pupil	2.6	4.8	12.7	6.7	13.7	6.1	9.7	7.2	10.4***	12.22***	.97
Response	32	Summarizes pupil's ideas or movement patterns, asks a pupil to demonstrate	0.6	1.5	1.4	1.6	2.0	2.5	1.4	2.0	3.26**	4.08***	1.58
Response	33	Compares the ideas or movement patterns expressed by one pupil to those of another or to these given, repeats pupil's ideas, asks a pupil to demonstrate	0.1	0.4	1.4	2.3	1.9	2.8	1.1	2.2	4.73***	5.47***	1.27
Initiation		Asks questions, initiates, terminates activity:											
Initiation	41	Asks questions requiring narrow answers, initiates short-term activity, terminates activity	12.6	5.6	8.6	4.3	8.5	3.8	9.9	5.0	-4.83**	-5.20**	-.20
Initiation	42	Broad, open questions which clearly permit choice in ways of answering and moving	0.9	1.4	1.9	1.4	1.9	1.3	1.6	1.5	4.55***	4.57***	-.16
Initiation		Content emphasis:											
Initiation	51	Presents information, opinions, demonstrates movement patterns, makes a pupil demonstrate	38.9	11.4	19.4	8.2	17.1	7.2	25.1	13.3	-11.98***	-13.98***	-1.79
Initiation	52	Organizes pupils, material, division of labour and responsibility	6.7	5.0	5.1	3.4	4.4	3.0	5.4	4.0	-2.21**	-3.39***	-1.39
Initiation	06	Gives directions, commands during activity (pupils expected to comply)	8.2	8.3	2.9	3.4	3.2	3.4	4.8	6.0	-5.12***	-4.82***	.54
Initiation	07	Criticizes pupil behaviour, rejects movement patterns, justifies authority	1.0	1.9	0.9	1.1	1.0	1.4	0.9	1.5	-.20	.07	.33
PUPIL TALK	08	Pupil answers question made by the teacher	1.5	1.9	2.8	2.3	3.4	2.7	2.6	2.5	3.66***	4.98***	1.60
PUPIL TALK	09	Pupuil initiates speech, asks for instructions, expresses own ideas or movement patterns	2.1	3.2	1.6	1.6	1.6	1.7	1.7	2.3	-1.19	-1.18	-.00
SILENT TEACHER	10	Teacher follows pupil's activity, silent quidance	19.2	12.6	31.2	12.1	31.7	12.6	27.4	13.6	5.86**	-5.98**	.24
SILENT TEACHER	11	Teacher's silent participation in movement activity	0.0	0.0	0.0	0.0	0.01	1.1	0.0	0.1	.00	.99	1.00
SILENT TEACHER	12	Confused situation, uproar	0.7	2.7	0.1	0.3	0.0	0.1	0.2	1.6	-1.97	-2.11*	-1.15
	CLUSTER II PUPILS COLLECTIVE MOVEMENT BEHAVIOUR	1 Pupils collectively passive	50.4	14.9	48.9	12.5	48.2	12.2	49.2	12.2	-.65	-.97	-.34
		2 Pupils collectively active	49.6	14.9	51.1	12.5	51.8	12.2	50.8	13.2	.65	.97	.34

* = $p < .05$ ** = $p < .01$ *** = $p < .001$

APPENDIX 6.2.2 Means, standard deviations of classtime by categories of modified PEIAC/LH-75 of the six teaching model groups in first and second microlesson 1976 (n = 74), 148 microlessons

First microlesson	1 N=15		2 N=15		3 N=15		4 N=15		5 N=9		6 N=5		N=74	
No	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
Cluster I														
1.	6.6	4.3	3.6	4.2	4.2	3.2	5.3	4.4	6.1	3.2	5.1	3.2	5.1	3.9
2.	2.2	2.2	4.0	2.9	3.8	3.1	9.0	7.1	6.7	4.1	6.5	3.1	5.1	4.7
3.1.	13.5	7.5	13.9	6.6	14.7	6.4	8.4	5.5	11.8	6.4	14.1	6.9	12.7	6.7
3.2.	1.5	2.1	2.6	2.1	1.3	1.1	0.7	1.0	1.1	0.8	1.0	1.4	1.4	1.6
3.3.	0.5	0.9	1.1	1.2	3.8	3.7	0.3	0.6	1.2	1.7	1.0	1.2	1.4	2.3
4.1.	11.0	4.5	9.4	3.5	6.1	3.4	8.5	4.8	7.4	4.7	8.7	2.0	8.6	4.3
4.2.	2.0	1.6	2.4	1.7	2.6	1.1	1.3	1.1	1.5	0.9	1.0	1.0	1.9	1.4
5.1.	20.3	6.4	18.5	7.1	16.5	5.6	21.8	10.6	19.2	10.8	20.8	10.4	19.4	8.2
5.2.	6.4	3.6	3.4	2.9	3.9	3.0	5.9	3.4	5.3	3.5	6.9	2.8	5.1	3.4
6.	2.8	4.3	2.0	2.0	1.2	2.2	4.0	4.0	4.8	2.4	4.2	3.5	2.9	3.4
7.	0.9	0.9	0.5	0.9	0.6	0.7	1.0	1.6	1.5	1.3	1.4	0.6	0.9	1.1
8.	1.8	1.5	3.3	2.5	2.0	1.8	2.6	2.2	3.7	2.5	5.3	2.5	2.8	2.3
9.	2.3	1.9	1.5	1.7	0.7	1.2	1.2	1.3	2.7	1.6	1.0	1.2	1.6	1.6
10.	28.1	11.6	33.6	11.6	38.3	10.7	29.9	13.1	27.0	12.4	23.0	6.2	31.2	12.1
11.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12.	0.0	0.0	0.0	0.0	0.1	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3
Cluster II														
1.	45.9	12.3	53.5	10.3	49.6	11.2	43.2	15.9	51.3	6.3	54.7	16.2	48.9	12.5
2.	54.1	12.3	46.5	10.3	50.4	11.1	56.8	15.9	48.7	6.3	45.3	16.2	51.1	12.5

Second microlesson	1 N=15		2 N=15		3 N=15		4 N=15		5 N=9		6 N=5		N=74	
No	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
Cluster I														
1.	6.2	4.3	3.8	2.2	4.0	2.2	3.8	3.9	6.5	4.2	3.6	1.2	4.6	3.4
2.	3.5	3.1	3.5	2.6	3.2	3.5	7.7	6.3	8.6	7.1	6.3	1.3	5.1	4.8
3.1.	16.7	6.2	15.0	7.0	12.2	4.3	10.6	7.1	14.3	3.9	13.3	3.9	13.7	6.1
3.2.	3.0	3.7	3.8	2.3	0.5	0.8	1.4	1.7	1.2	1.6	1.0	1.2	2.0	2.5
3.3.	1.3	1.6	0.9	1.1	5.7	3.7	0.2	0.4	1.7	2.3	0.6	0.6	1.9	2.8
4.1.	10.0	4.3	9.0	4.2	6.7	3.2	6.7	3.7	6.8	3.1	9.9	3.1	8.5	3.8
4.2.	2.0	0.9	2.8	1.7	2.2	1.1	1.2	1.0	1.2	1.0	1.0	0.7	1.9	1.3
5.1.	17.2	6.5	17.8	7.5	17.3	8.0	16.7	7.7	12.5	5.6	17.6	6.8	17.1	7.2
5.2.	5.3	2.5	2.5	1.7	3.3	2.5	5.5	3.8	4.9	3.1	6.7	2.9	4.4	3.0
6.	3.4	4.6	2.4	2.1	1.5	1.8	3.0	2.8	6.1	3.1	5.3	5.0	3.2	3.4
7.	0.6	0.7	0.8	1.7	0.5	0.8	1.4	1.6	1.9	1.6	1.0	1.4	1.0	1.4
8.	3.0	2.4	2.7	2.5	2.9	2.4	4.0	3.5	4.5	2.5	4.6	2.5	3.4	2.7
9.	1.6	1.1	1.0	1.4	0.9	1.5	1.8	2.1	3.3	2.3	1.6	1.4	1.6	1.7
10.	26.1	12.1	34.0	11.5	39.1	8.9	32.0	15.5	26.5	13.0	27.5	5.8	31.7	12.6
11.	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.0	0.0	0.0	0.0	0.0	0.1
12.	0.1	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
Cluster II														
1.	48.0	11.8	50.0	9.2	49.6	11.3	45.6	13.5	48.0	14.6	47.3	19.4	48.2	12.2
2.	52.0	11.8	50.0	9.2	50.4	11.3	54.4	13.5	52.0	14.6	52.7	19.4	51.8	12.2

APPENDIX 6.3.1 Means, standard deviations, and T-tests for the PEIAC/LH-75 Ii categories scores across the three micro-teaching lessons for the male students by using a Multiple Range, Scheffé Procedure *)

Variable categories		Lesson 1. (control)		Lesson 2.		Lesson 3.		Total (n=63)		Pairs of lessons groups significantly different at the 0.05 level and p
Cluster	No	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
I	1	1.0	(2.0)	1.8	(1.7)	2.5	(2.1)	1.8	(2.0)	
	2	5.5	(3.6)	8.5	(3.6)	6.3	(3.1)	6.8	(3.6)	1 - 2*
										2 - 3*
	3.1	2.9	(3.8)	6.3	(4.8)	6.6	(4.1)	5.3	(4.5)	2 - 3*
										1 - 2*
	3.2	0.3	(1.0)	3.1	(3.0)	2.4	(2.5)	2.9	(2.6)	1 - 3**
										1 - 2***
	3.3	0.9	(2.0)	1.4	(2.8)	2.1	(3.4)	1.4	(2.8)	1 - 3***
	4.1	14.2	(6.3)	13.3	(4.5)	11.9	(5.1)	13.1	(5.5)	
	4.2	0.4	(0.8)	1.7	(1.5)	1.2	(1.4)	1.1	(1.4)	1 - 2**
										1 - 3*
	5.1	38.6	(11.5)	23.1	(7.1)	22.5	(6.0)	28.1	(11.3)	1 - 2***
										1 - 3*
	5.2	12.5	(7.1)	5.8	(2.8)	5.9	(3.1)	8.0	(5.7)	1 - 2**
										1 - 3***
	6	3.5	(8.8)	5.5	(8.8)	1.8	(3.7)	3.6	(7.5)	2 - 3*
	7	0.3	(1.0)	1.0	(1.5)	1.4	(1.5)	0.9	(1.4)	1 - 2*
	8	1.5	(2.8)	2.3	(2.3)	2.4	(2.5)	2.1	(2.5)	
	9	0.9	(1.5)	1.7	(1.7)	1.5	(1.3)	1.4	(1.5)	1 - 2*
	10	15.9	(11.1)	24.1	(9.3)	31.0	(10.7)	23.7	(12.0)	1 - 2**
	11	1.2	(3.5)	0.4	(2.0)	0.1	(0.2)	0.6	(2.3)	2 - 3***
	12	0.5	(2.2)	0.0	(0.0)	0.3	(1.3)	0.3	(1.5)	
Cluster II	1	57.9	(10.5)	53.0	(11.8)	51.1	(8.7)	54.0	(10.6)	1 - 3*
										2 - 3*
	2	42.1	(10.5)	47.0	(11.8)	48.9	(8.7)	46.0	(10.6)	1 - 3*
										2 - 3*

(T-values were calculated after applying Barlett's (1937) test for homogeneity of variance)
*) Scheffé (1959)

* = significant at the 5% level
** = signigicant at the 1% level
*** = significant at the 0.1% level

APPENDIX 6.3.2 Means, standard deviations, and T-tests for the PEIAC/LH-75 Ii categories scores across the three micro-teaching lessons for the female students by using a Multiple Range, Scheffé Procedure *)

Variable categories		Lesson 1. (control)		Lesson 2.		Lesson 3.		Total (n=63)		Pairs of lessons groups significantly different at the 0.05 level and p
Cluster	No	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
I	1	1.6	(2.0)	2.1	(2.2)	2.2	(2.3)	2.0	(2.1)	1 - 2**
	2	5.0	(2.6)	8.6	(5.8)	10.8	(5.5)	8.1	(5.3)	1 - 3***
	3.1	0.7	(1.3)	4.8	(3.7)	4.6	(3.2)	3.4	(3.5)	1 - 2** 1 - 3***
	3.2	0.2	(0.6)	2.0	(2.8)	2.3	(2.2)	1.5	(2.2)	1 - 2** 1 - 3***
	3.3	0.1	(0.4)	1.1	(2.1)	1.6	(3.0)	0.9	(2.2)	1 - 2* 1 - 3*
	4.1	14.3	(5.3)	12.8	(4.0)	12.3	(4.0)	13.1	(4.5)	
	4.2	0.1	(0.4)	1.2	(1.4)	1.4	(1.2)	0.9	(1.2)	1 - 2** 1 - 3***
	5.1	42.2	(8.5)	27.9	(6.5)	21.9	(4.6)	30.7	(10.8)	1 - 2*** 1 - 3*** 2 - 3***
	5.2	9.1	(4.7)	5.9	(4.0)	5.0	(3.4)	6.7	(4.4)	1 - 2* 1 - 3**
	6	2.4	(3.9)	5.4	(8.4)	4.3	(9.0)	4.0	(7.4)	
	7	0.2	(0.6)	0.9	(1.2)	0.6	(0.9)	5.5	(0.9)	1 - 2*
	8	0.6	(1.1)	2.4	(2.1)	2.4	(1.8)	1.8	(1.9)	1 - 2** 1 - 3**
	9	1.2	(1.5)	1.8	(1.6)	1.9	(1.7)	1.6	(1.6)	
	10	21.0	(9.3)	22.9	(9.5)	27.8	(8.3)	23.9	(9.4)	1 - 3* 2 - 3*
	11	0.2	(3.3)	0.1	(0.6)	0.1	(0.2)	0.7	(2.2)	
	12	0.2	(0.9)	0.1	(0.2)	0.1	(0.5)	0.1	(0.6)	
Cluster II	1	55.6	(1.2)	56.3	(7.5)	46.0	(6.3)	52.7	(9.7)	1 - 3** 2 - 3***
	2	44.4	(11.2)	43.7	(7.5)	54.0	(6.3)	47.3	(9.7)	1 - 3** 2 - 3***

(T-values were calculated after applying Barlett's (1937) test for homogeneity of variance)
*) Scheffé (1959)

* = significant at the 5% level
** = signigicant at the 1% level
*** = significant at the 0.1% level

APPENDIX 6.3.3 Significance of the differences between factor scores estimated for the 1st (cont.), 2nd and 3rd microlesson (n = 74), 221 lessons, analysis of variance and t-test (ANOVA)

Factors No	1st (cont.) n=73 Mean	SD	2nd micro-lesson n=74 Mean	SD	3rd micro-lesson n=74 Mean	SD	Diff 1-2 df=146 t	Diff 1-3 df=146 t	Diff 2-3 df=146 t	Diff df=2 df=21 t
1. Teacher initiation (-) vs. teacher response behavior (+)	408	.74	538	.73	554	.81	10.8***	11.4***	1.23	82.15***
2. Channel of teacher - pupil communication verbal (-) vs. motor (+)	497	.96	500	.96	503	1.09	.20	.36	.17	.94
3. Teacher silence (-) vs. teacher feedback and motivational communication	457	.96	515	.89	528	1.02	3.79**	4.34**	.81	11.38***

** = p<0.01
*** = p<0.001

APPENDIX 6.3.4 Means standard deviations, and t-test for the Indices appearing in connection with PLAC/LH-75 II system across the three microteaching lessons for intake course 1986/1988 male students (n = 21) by using Multiple Range Test, Schéffe procedure *) and analysis of variance

Variables Indices	Lesson 1. (control) Mean	SD	Lesson 2. Mean	SD	Lesson 3. Mean	SD	Total Mean	SD	Pairs of lesson groups significantly different at the 0.05 level and p
1. Percent teacher talk (TT)	80.0	(10.6)	71.4	(8.6)	64.8	(9.9)	72.1	(11.5)	1 - 2** 1 - 3*** 2 - 3**
2. Percent pupil Talk (PT)	2.4	(3.1)	4.1	(2.7)	3.1	(3.1)	3.4	(3.0)	1 - 2**
4. Teacher silent duidance and participation in movement activity ratio (TSQPR)	18.1	(10.3)	25.5	(8.9)	32.2	(10.8)	25.4	(11.5)	1 - 2** 1 - 3*** 2 - 3**
5. Teacher response ratio (TRR)	81.1	(33.4)	78.2	(25.1)	86.2	(15.6)	81.9	(25.5)	
6. Corrected teacher response behavior ratio (ID-index)	33.1	(12.6)	51.5	(13.1)	51.3	(8.7)	45.3	(14.4)	1 - 2*** 1 - 3***
7. Content emphasis ratio (CCR)	65.6	(12.8)	43.8	(7.6)	41.5	(7.8)	50.3	(14.5)	1 - 2*** 1 - 3***

(T-values were calculated after applying Barlett's (1937) test for homogeneity of variance)
*) Scheffé (1959)
* = significant at the 5% level
** = significant at the 1% level
*** = significant at the 0.1% level

APPENDIX 6.3.5 Means standard deviations, and t-test for the Indices appearing in connection with PEIAC/LH-75 II system across the three microteaching lessons for intake course 1986/1988 female students (n = 21) by using Multiple Range Test, Schéffe procedure *) and analysis of variance

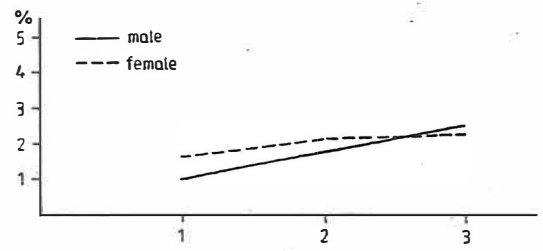
Variables Indices significantly	Lesson 1. (control)		Lesson 2.		Lesson 3.		Total		Pairs of lesson groups different at the 0.05 level and p
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
1. Percent teacher talk (TT)	75.8	(8.8)	72.8	(9.3)	67.1	(7.6)	71.9	(9.2)	1 - 3** 2 - 3**
2. Percent pupil Talk (PT)	1.8	(1.8)	4.1	(2.0)	4.3	(2.5)	3.4	(2.4)	1 - 2** 1 - 3
4. Teacher silent duidance and participation in movement activity ratio (TSQPR)	22.8	(8.7)	24.0	(9.8)	29.7	(8.2)	25.0	(9.3)	1 - 3* 2 - 3**
5. Teacher response ratio (TRR)	81.7	(29.5)	77.5	(24.7)	85.0	(20.6)	81.4	(25.0)	2 - 3*
6. Corrected teacher response behavior ratio (ID-index)	29.8	(7.7)	45.5	(14.2)	53.3	(11.8)	42.9	(15.0)	1 - 2*** 1 - 3*** 2 - 3***
7. Content emphasis ratio (CCR)	65.7	(8.7)	47.9	(8.8)	40.6	(8.2)	42.9	(13.5)	1 - 2*** 1 - 3*** 2 - 3***

(T-values were calculated after applying Barlett's (1937) test for homogeneity of variance)
*) Scheffé (1959)

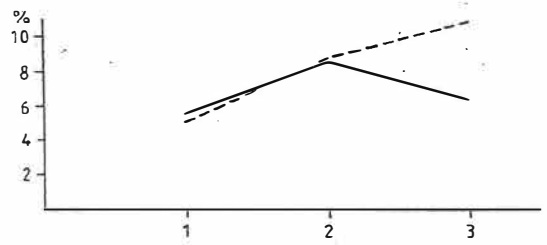
* = significant at the 5% level
** = significant at the 1% level
*** = significant at the 0.1% level

COMPARISON OF STUDENTS PROCESS BEHAVIOR IN MICROLESSONS (COURSE 1988) 1. (CONTROL) 2. AND 3. BY CATEGORY AND INDEX.

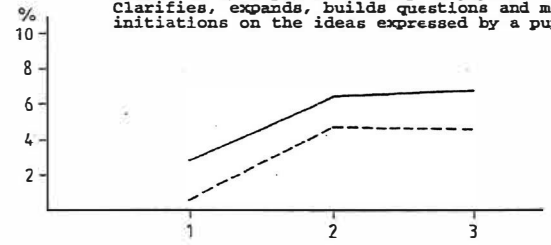
Cat. 1. Praises, encourages, accepts the feeling tone of a pupil



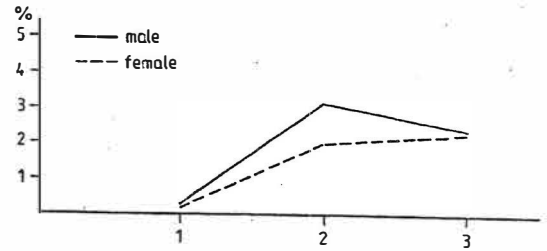
Cat. 2. Gives corrective feedback, directs, clarifies, answers pupil's questions



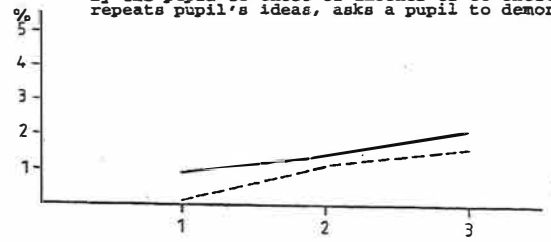
Cat. 3.1. Make use of the ideas and movement patterns suggested by a pupil or group of pupils. Clarifies, expands, builds questions and movement initiations on the ideas expressed by a pupil.



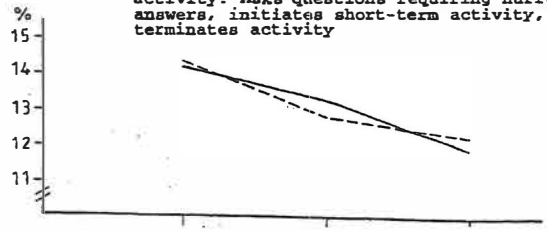
Cat. 3.2. Summarizes pupil's ideas or movement patterns, asks a pupil to demonstrate



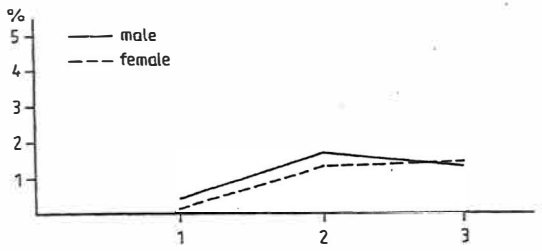
Cat. 3.3. Compares the ideas or movement patterns expressed by one pupil to those of another or to those given, repeats pupil's ideas, asks a pupil to demonstrate



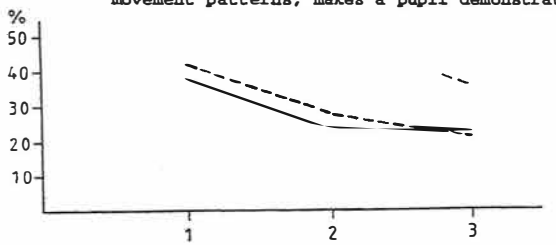
Cat. 4.1. Asks questins, initiates, terminates activity: Asks questions requiring narrow answers, initiates short-term activity, terminates activity



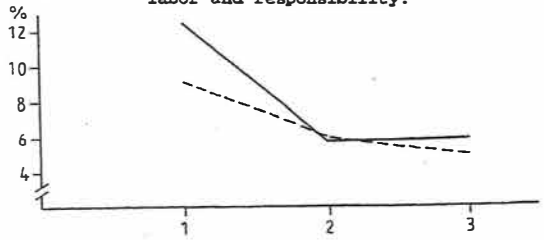
Cat. 4.2. Broad, open questions which clearly permit choice in ways of answering and moving



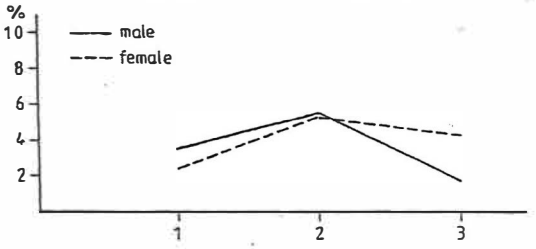
Cat. 5.1. Present informations, opinions, demonstates movement patterns, makes a pupil demonstrate



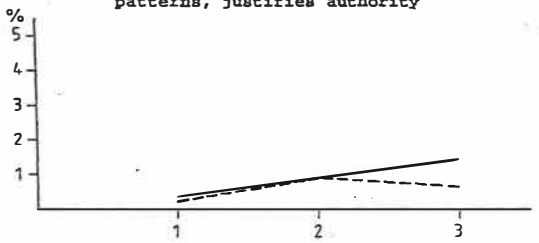
Cat. 5.2. Organizes pupils, material, division of labor and responsibility.



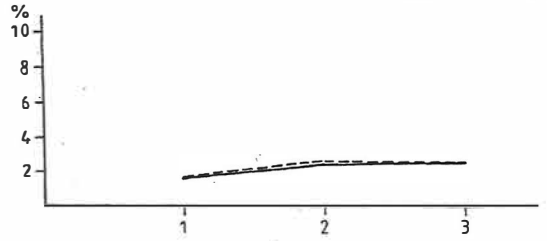
Cat. 6. Gives directions, commands during activity (pupils expected to comply)



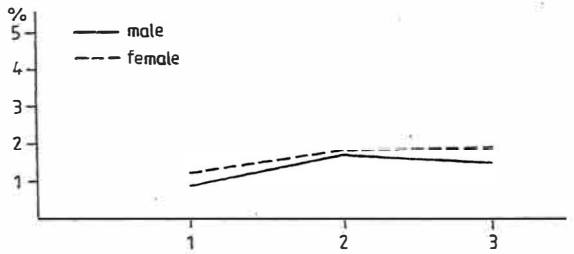
Cat. 7. Criticizes pupil behavior, rejects movement patterns, justifies authority



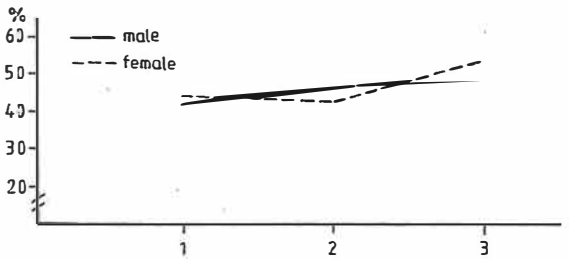
Cat. 8. Pupil answers questions made by the teacher



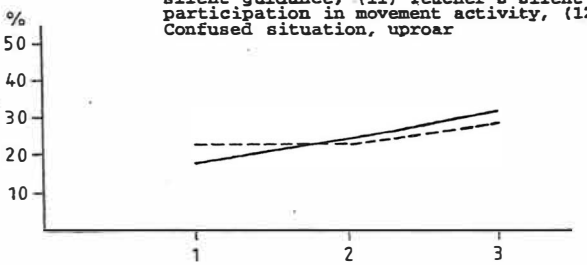
Cat. 9. Pupil initiates speech, asks for instruction, expresses own ideas or movement patterns



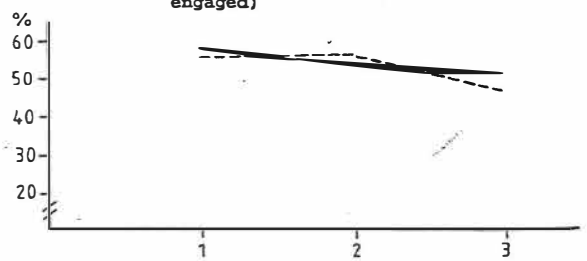
Cluster 2. Pupils collectively active



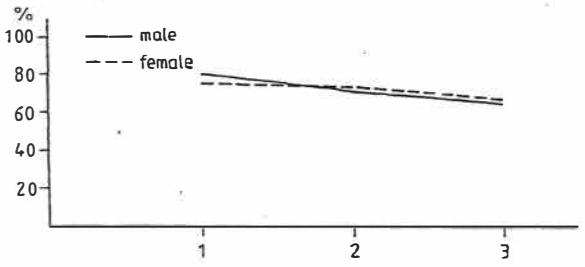
Cat. 10-12. (10) Teacher follows pupil's activity, silent guidance, (11) Teacher's silent participation in movement activity, (12) Confused situation, uproar



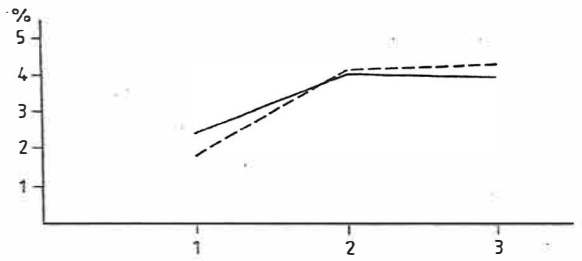
Cat. 1. Pupils collectively passive (not motor engaged)



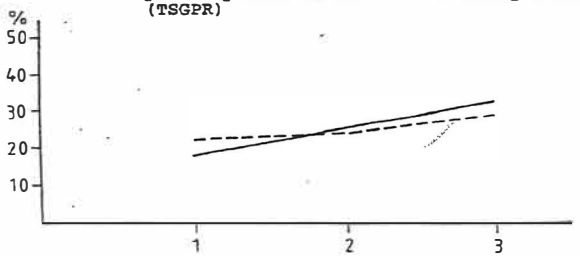
Index 1. Percent teacher talk (TT)



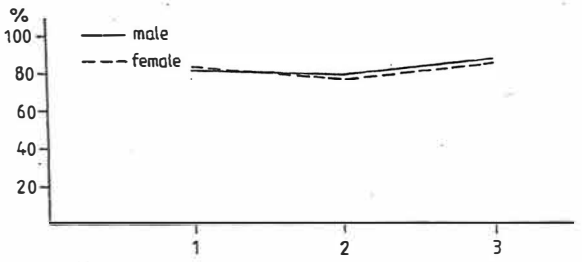
Index 2. Percent pupil talk (PT)



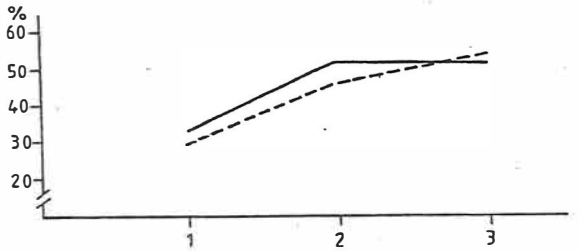
Index 4. Teacher's silent guidance and silent participation in movement activity ratio (TSGPR)



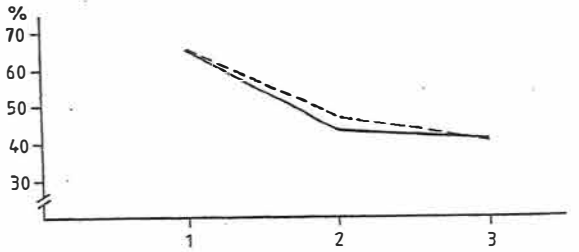
Index 5. Teacher response ratio (TRR)



Index 6. Corrected teacher response behavior ratio (ID-index)



Index 7. Content emphasis ratio (CCR)



APPENDIX 6.4.1 Pearson's correlation coefficients between PEIAC/LH-75 II categorie scores across the three micro lessons (n = 126) the highest correlation coefficient on the diagonal

Variables Cluster	Categ. No.	1.	2	3.1	3.2	3.3	4.1	4.2	5.1	5.2	6	7	8	9	10	11	12	1	2
I	1.	.46																	
	2.	.46	.46																
	3.1	.12	.06	.54															
	3.2	.02	-.04	.54	.54														
	3.3	.21	.04	.29	.20	-.36													
	4.1	-.24	-.15	-.12	-.07	-.36	-.36												
	4.2	.15	.05	.45	.45	.32	-.13	.45											
	5.1	-.24	-.29	-.52	-.54	-.20	.03	-.40	-.54										
	5.2	-.09	-.18	-.31	-.29	-.22	-.03	-.18	.30	.37									
	6.	-.26	.03	-.23	-.13	-.11	.07	-.21	.02	-.26	-.39								
	7.	.03	.10	.12	.07	.14	.04	.01	-.26	-.10	-.14	-.26							
	8.	.14	.12	.32	.02	.26	.15	.22	-.36	-.22	-.09	.21	-.36						
	9.	-.13	.16	-.10	.04	-.11	.04	-.06	-.11	-.08	.11	.03	-.00	.16					
	10.	.10	.09	.19	.30	.04	-.32	.16	-.54	-.20	-.39	.12	-.05	-.05	-.54				
	11.	-.13	-.16	-.12	-.07	-.13	.17	.03	.20	.16	-.07	-.12	.05	.13	-.33	-.33			
	12.	-.03	.01	-.08	.00	.13	.10	-.08	.06	.10	.22	.02	.02	-.16	-.05	-.05	.22		
II	1.	-.24	-.33	-.16	-.14	-.08	.21	-.15	.21	.37	.11	.03	.04	.03	.28	.04	.07	-.100	
	2.	.24	.33	.16	.14	.08	-.21	.15	-.21	-.37	-.11	-.03	-.04	-.03	.28	-.04	-.07	-1.00	-100

Determination of the correlation matrix = 0013610

APPENDIX 6.4.2 Principal component analysis on students' process behavior (PEIAC/LH-75 II) variables across three successive microlessons, (n = 126; n = 42)

Cluster, Categories No	Factor loadings 1.	2.	3.	h^2
I				
Teachers initiations (-) vs response behavior (+)				
5.1. Present information, opinions, demonstrates movement patterns, makes a pupil demonstrate	-.78	-.23	.00	.66
3.1. Makes use of the ideas and movement patterns suggested by a pupil: clarifies, expands, builds questions and movement initiations on the ideas expressed by a pupil	.70	.04	.16	.52
3.2. Summarizes pupil's ideas or movement patterns, asks a pupil to demonstrate	.68	-.09	.15	.50
4.2. Makes questions requiring higher level of thinking or activity	.54	.07	.22	.34
5.2. Organizes pupils, material, division of labour and responsibility	-.49	-.18	.29	.36
8. Pupil answers question made by the teacher	-.36	.09	-.12	.15
3.3. Compares the ideas or movement patterns expressed by one pupil to those of another or to those given, repeats pupil's ideas, asks a pupil to demonstrate	.33	.16	.17	.17
7. Criticizes pupil behavior, rejects movement	.21	.06	.01	.05
Teacher motivational feedback during pupils' collective activity (+)				
2. Gives corrective feedback, directs, clarifies, answers pupil's questions	.06	.80	-.25	.71
1. Praises, encourages, accepts the feeling tone of a pupil	.10	.61	.20	.42
II/2 Pupil's collective movement activity/passivity	.21	.41	.07	.22
Teacher gives direction during pupils' collective activity (-) – vs. silence (+)				
6. Gives directions, commands during activity (pupils expected to comply)	-.09	-.14	-.62	.41
10-12 (10) Teacher follows pupil's activity, silent guidance (11) Teacher's silent participation in movement activity (12) Confused situation, uproar	-.36	.10	.42	.32
4.1. Asks questions requiring narrow answers, initiates short-terms activity, terminates activity	-.06	-.29	-.32	.19
9. Pupil initiates speech, asks for instruction, expresses own ideas of movements	.02	.04	-.23	.05
Eigenvalue	3.0	1.2	0.9	5.1
% common variance	58.8	22.8	18.3	100
% total variance	19.9	7.7	6.2	33.8

(determinant of corr. matrix .0013610, $p<0.05$)

APPENDIX 6.4.3 Means, standard deviations, and T-tests for the three process behavior factors scores across the three micro-teaching lessons (n = 126) by using a Multiple Range test, Scheffé Procedure *)

Varimax factors	Lesson 1. (n=42)		Lesson 2. (n=42)		Lesson 3. (n=42)		Total (n=126)		Pairs of lessons significantly different at the 0.05 level
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
1. Teacher initiation (-) vs. response behavior (+) (19.9%)	-.86	(0.8)	.32	(0.9)	.54	(0.6)	.00	(1.0)	1 - 2 1 - 3
2. Teacher motivational feedback during activity (+) - vs. passivity (-) (7.7%	-.41	(0.8)	.13	(0.9)	.28	(0.9)	.00	(1.0)	1 - 2 1 - 3
3. Teacher silence (+) vs. gives direction during pupil's collective activity (-) (6.2%)	.14	(0.7)	-.21	(0.9)	.08	(0.8)	.00	(0.8)	

(T-values were calculated after applying Barlett's (1937) test for homogeneity of variance)
*) Scheffé (1959)

APPENDIX 7 **Student's entry teaching behavior**
APPENDIX 7.1 Teaching episode rating scale

SELECTION TEST 1976 - 1988

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF PHYSICAL EDUCATION

TEACHING EPISODE

You will have 7 minutes to get acquainted with the following teaching task exercise, and with the available space and equipment. After that you will teach the task to your own group and imagine that they are your pupils at school. You will have to bring about ability. The teaching episode lasts 3.5 minutes. It is divided into two parts:
1) given teaching task (2 min)
2)its development and modification (1.5 min)

Judges will pay attention to how *clearly* you are able to present to the "pupils" the objectives, content and activity instruction of the given task and how aptly you are able to use your own imagination and judgement in developing the task and - and how you interact with your pupils.

You may use the text, if need be, but try to present the task freely and using your own words.

APPENDIX 7.2 Reliability of the rating scale to measure student's entry teaching skills'; inter-rater agreement and stability by using Kendall's coefficient of concordance (W), Chi Square determined correlation coefficients, r^n

TEST 1

Item	Ratings: 1. W	r^2	chi^2 df=74 df=1 p		2. W	r^2	chi^2	df=1 p
1. Presentation	.73	.46	30.1	.00	.94	.88	16.9	.05
2. Understanding of task content	.76	.51	42.1	.00	.91	.82	16.3	.06
3. Teacher - pupil interaction	.74	.49	41.1	.00	.76	.53	13.8	.13
4. Creativity	.76	.51	43.1	.00	.86	.73	15.6	.08

TEST 2

Item	Ratings: 1. W	r^2	chi^2 df=41 p		2. W	r^2	chi^2	df=1 *)41 p
1. Presentation	.70	.60	57.6	.04	.50	.34	20.8	.01
2. Understanding of task content	.56	.30	45.7	.28	.18	.22	7.8	.01
3. Teacher - pupil interaction	.69	.60	56.9	.05	.15	.32	6.3	.01
4. Creativity	.65	.59	53.6	.01	.34	.25	14.2	.00
5. Sum scores	.68	.58	55.6	.06	.26	.36	76.1	.00

Two rathers; unit: videorecorded 5 min micro lesson control ratings 1. and 2. between four weeks interval; *) missing cases =3,

APPENDIX 7.2.1 Microteaching course: Exercise 1

MICROTEACHING COURSE

Exercise 1: Teaching episode/5 min and rating of teaching skill (control)

Instructions:

1) You will get information about the main objective (random selection, 1-15).
2) You will be informed about the sub-domains of subject matter from which you can choose the one you prefer (ball games, rhythmic movement expression, apparatus gymnastics, folk dancing, play).
3) You will be free to choose the teaching task and the teaching form. For instance, you may choose ball games and teach taking the ball forward, as pair work, etc.
4) You have 10 min to prepare.

You will teach your own practice group (5-10 students) for 5 minutes and imagine that they are your pupils at school

SELECTION TEST 1976-1988

Teaching episode
Rating of teaching skill

Student's group		Rating dimensions			Results	
	Presentation - voice quality - expression - fluency - clarity - movement	Understanding of task content - main points - phases - instruction	Communication - interaction with pupils - directivenes of main points - obsrevation - feedback	Creativity - originality - aptness - presentation of main points	Points	Rank
N1 Name	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6		

APPENDIX 7.3 Correlations of students' entry teaching skills' ratings between (intake and control teaching episode) sum scores in four microteaching course populations (n = 205)

Variables	Intake course
	1974/1976 n=69
	(1) (2)
Teaching episode 1	-
Teaching episode 2	x).30* -
	1976/1979 n=53
	(1) (2)
Teaching episode 1	-
Teaching episode 2	.23* -
	1977/1980 n=41
	(1) (2)
Teaching episode 1	-
Teaching episode 2	.26* -
	1986/1988 n=42
	(1) (2)
Teaching episode 1	-
Teaching episode 2	.52** -

(4 items, two observers)
*=significant at the 5% level; **=significant at the 1% level
x)= measured after students intake

APPENDIX 8 **Students attitudes: "ideal" P.E. teacher expectation**
APPENDIX 8.1 Questionnaire for students' "ideal" P.E. teacher expectation study

Teacher Expectations Questionnaire _____
Number

Name: _____

IDEAL TEACHER

Let us imagine that you will start as a student at some school. You are given the rare opportunity to choose your own teacher before the beginning of view.

On the next few pages you will find a number of characteristics describing teachers and your task is to place your ideal teacher on the scale with regard to each characteristics. Only the opposite extremes of each characteristics are given and, to make the assessment easy, each characteristics is explained in a few words.

You make assessment by drawing a circle round the number in the scale which, according to your opinion, describes your ideal teacher best. If your ideal teacher has a lot of the characteristics described at the left extreme you draw a circle round number 6. If, on the other hand, your ideal teacher has a lot of the characteristics described at the right extreme of the scale you draw a circle round number 1. Numbers 5 and 2 indicate that your ideal teacher has, to some extent, the characteristics described at the left or right extreme of the scale. Numbers 4 and 3 are in the middle of the scale so if you place a circle round either of them your ideal teachers has only some of the characteristics described at the left or right extreme. Try to avoid, however, using too much of the middle values and do not hesitate to use numbers 6 and 1.

Notice that we are interested in your current ideal P.E. teacher. Do not let your previous ideals influence your assessment. Before handing in the questionnarie check that you have written your name on it and assessed all the features. Make sure that you have drawn only one circle on each line.

1. systematic	6 5 4 3 2 1	responsive
- the lessons have a clear outline or disposition which is followed		- the lessons develop on the basis of the subject or topic
2. sure in his/her opinions	6 5 4 3 2 1	unsure in his/her opinions
- gives his/her opinion on various subjects quickly: sticks to it		- does not readily give his/ opinion; changes his/her opinion
3. sociable	6 5 4 3 2 1	aloof
- participates in the leisure activities, hobbies, etc. of the students		- does not participate in the activities outside the classroom

4. friendly	6 5 4 3 2 1	rude
- digresses from the subject to be pleasant and/or to help his/her students		- points out the mistakes of the students without paying attention to their feelings; overcritical
5. serious	6 5 4 3 2 1	humorous
- instruction is stricktly matter-of-fact		- uses humor as an aid in instruction
6. reviews subject matter	6 5 4 3 2 1	constantly presents new subject matter
- repeats regularly topics dealt with during the previous lessons		- never repeats topics dealt with during the previous lessons
7. fact-centered	6 5 4 3 2 1	student-centered
- pays attentions only to the academic achievements of the students; personal problems		- takes the students' personal problems and needs always into account
8. gets involwed	6 5 4 3 2 1	placid
- may, in the middle of a lessons, start to explain volubly an issue related to the topic dealt with		- does not digress from the subject
9. conventional	6 5 4 3 2 1	willing to experiment
- uses old and recognized teaching methods		- prepared to use new teaching methods and equipment
10. theoretical	6 5 4 3 2 1	practical
- teaching is objective and based on critical analysis of facts		- teaching is firmly linked up with practical applications
11. activates the students	6 5 4 3 2 1	gives the facts
- concentrates on activating the students; tries to get everyone to work		- concentrates on giving new facts; follow-up of the students' work secondary
12. narrow-ranging	6 5 4 3 2 1	wide-ranging
- sticks to his/her own subject in teaching; tries to find the examples within the subject		- does not limit his/her teaching to his/her own subject; takes the examples from various sources and other subjects

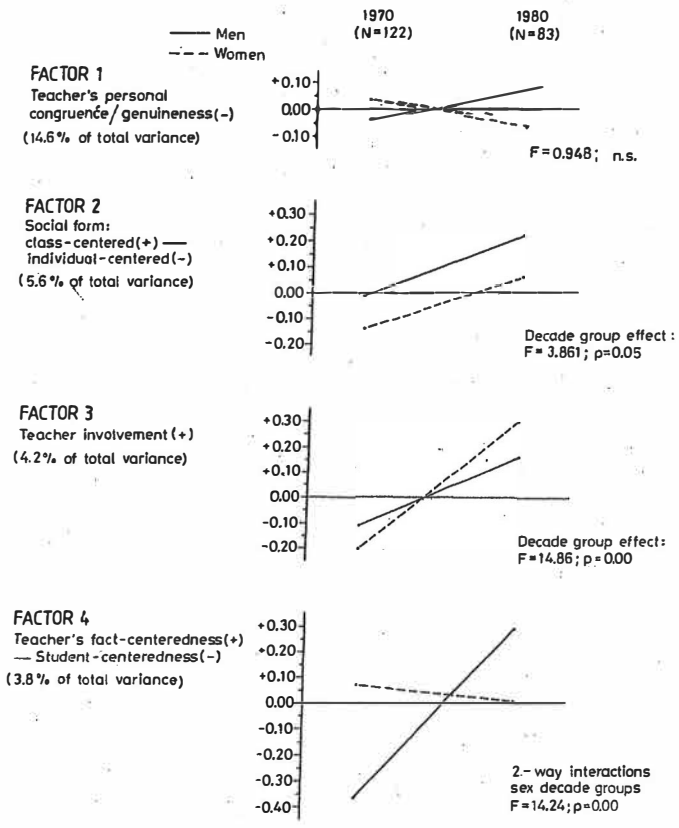
13. distant	6 5 4 3 2 1	approachable
- stiff and formal in his/her relations with the students		- accessible to all students; addresses his /her students as his/her equals
14. responsibility	6 5 4 3 2 1	evades responsibility
- willing to take responsibility over decisions concerning the class		- avoids decisions concerning the class
15. strict in keeping discipline	6 5 4 3 2 1	slack in keeping discipline
- the teacher has detailed rules and directions for various situations		- the teacher has no present rules
16. easy-going	6 5 4 3 2 1	matter-of-fact
- likes to tell about him/herself and his/her own experiences		- leaves his/her own personality outside the teaching; sticks to facts
17. encouraging	6 5 4 3 2 1	supervising
- guides purposefully the practical exercises of the whole class by encouraging and keeping the initiative		- guides when necessary: encourages less verbally: supervises
18. class-centered	6 5 4 3 2 1	individual-centered
- concentrated on the guidance and observation of the activities of the whole class		- spends time on encouraging individuals; strengthens personal expression, individual performance
19. directs expression	6 5 4 3 2 1	expressive
- guides the students' P.E. expression directing it according to his/her own preferences		- identifies him/herself with the activities of the students and participates in them, expresses his/her similar emotions spontaneously
20. bears responsibility alone	6 5 4 3 2 1	delegates responsibility
- plans the activities of the group in advance and supervises the attainment of the goals		- gives the students the opportunity to choose suitable role tasks in group activities, avoids responsibility in planning

APPENDIX 8.2 Significant ($p < .05$) correlations between selected "ideal" P.E. teacher expectations questionnaire items *)

	1.	3.	5.	7.	8.	9.	12.	13.	14.	15.	16.	17.	18.	19.
1.	-													
3.		-												
5.		-21	-											
7.		-29	19	-										
8.					-									
9.		-22	27	25	19	-								
12.			14	23	-20	34	-							
13.		-17	27	13		22	18	-						
14.	14				24	-15	-22	15	-					
15.	23			16			14			-				
16.		17	20	-21	28	-13	-13	-24	27		-			
17.		14	-13	13								-		
18.				24		23	21	25			-15	18	-	
19.		14				18	19	23					32	-

*) Determinant of correlation matrix .1942928

APPENDIX 8.3 The average location of decade 1970's and 1980's course groups (n = 205) in factor structure dimensions of "ideal" P.E. teacher expectations based on their means and standard deviations



APPENDIX 8.4 Summary of regression analyses for male and female groups: % of variance of success in student teaching (C_3 = theory test scores, C_4 = practice test scores, C_5 = final mark) explained by students' attitudes (= personal expectations concerning "ideal" P.E. teacher characteristics)

Year/course group	Males			Females		
	C_3	C_4	C_5	C_3	C_4	C_5
1976	-	-	F4:17 %	-	F2:6,5 % F3:12,8 %	-
1979	-	F4:16 %	F1:16 %	F2:7 %	-	F2:8 %
1980	-	F2:19,5 %	-	F4:10 %	-	-
1988	-	F1:34 %	F1:29 %	-	-	-

C_1= F1: Measured teaching behavior ("teacher initation (-) or response behavior (+)
C_2= Corrected ID-index

	Males		Females	
	C_1	C_2	C_1	C_2
1988	F:16 %	F1:12 %	-	-

APPENDIX **9** **Program predictive validation:**
APPENDIX 9.1 Procedures used in the selection of future P.E. teachers

The two stages and tests have had different weights in the final selection procedure in different years as follows:

	Year 1974	Year 1976	Year 1977	Year 1979	Year 1986
Stage I (prior school)	50 %	25 %	25 %	20 %	-
Stage II					
-theory test	-	-	-	15 %	30.7 %
-practice test	50 %	50 %	50 %	45 %	55.5 %
-oral test or micro-teaching episode	*	25 %	25 %	15 %	11.5 %
- musical test	-	**	**	**	2.3 %
- psychological test	-	-	-	5 %	-
Total point	100	100	100	100	100

* 7 % + 43 % = 50 % practical test
** in stage I scores

APPENDIX 9.2 Data collection

APPENDIX 9.2.1 Students selection procedure phase one; the minimum points of students in intake by year and sex

Variables of intake phase one	1974 (n=76)		1976 (n=60)		1977 (n=50)		1979 (n=43)		1986 (n=47)	
	M(30)	F(24)	M(24)	F(36)	M(20)	F(22)	M(22)	F(21)	M(24)	F(23)
External matriculation examination	18	20	23	21	18	24	18	26	11	15
Previous school achievement in terms of secondary school learning report (+ sport weighted by 3)	7.2*	7.45*	8.30	9.30	7.45	7.50	7.60	8.10	6.5	6.9

(* mean of theory subjects)
M=male
F=female

APPENDIX 9.2.2 Data collection and drop out in four intake course population

	Intake – course group							
Student's intake population	1974		1976		1977		1986	
- year								
- gender	n=76		n=60		n=50		n=47	
(m) (f)	(30)	(46)	(24)	(36)	(20)	(30)	(24)	(23)
Students in course of didactic observation and microteaching	1976		1979		1980		1988	
- year								
- gender	n=69		n=53		n=41		n=42	
(m) (f)	(26)	(43)	(21)	(32)	(16)	(25)	(21)	(21)
1) drop out: % intake -> microteaching	13.3%	6.5%	12.5%	11.1%	20.0%	16.7%	12.4%	8.9%
2) students, who do not have examination in intake procedure, Stage I (school achievement)	9.2%		11.6%		18%		10.6%	

APPENDIX 9.3 Description of subpopulations:

APPENDIX 9.3.1 Performance of subjects in four student selection variables and weighted sum score: comparison by percent means and standard deviations between gender groups by analysis of variance (ANOVA) and t-test

	Intake-course 1974/76						
	Male (n=26)		Female (n=43)		Total (n=69)		2-tail t-test
Variables	x	sd	x	sd	x	sd	p
(1) Stage I: sum scores	46.2	6.1	46.7	6.6	46.5	6.4	
Stage II:							
(2) Theorytest scores	24.8	32.7	28.6	41.5	27.1	38.2	
(3) Practice test sum scores	37.3	29.8	38.2	24.1	37.9	26.1	
(4) Teaching episode *) sum score	15.4	3.1	13.8	5.6	14.4	4.9	
Total wighted							
(5*) sum score	64.8	22.8	66.3	25.8	65.8	24.5	
Missing cases 0							

	Intake-course 1976/79						
	Male (n=21)		Female (n=32)		Total (n=53)		2-tail t-test
Variables	x	sd	x	sd	x	sd	p
(1) Stage I: sum scores	553.0	71.1	505.8	55.7	524.5	65.9	**
Stage II:							
(2) Theorytest scores	128.1	25.6	118.2	18.8	122.1	22.0	
(3) Practice test sum scores	294.3	15.8	295.6	22.2	295.1	19.8	
(4) Teaching episode *) sum score	149.1	22.4	156.8	15.6	153.8	18.8	
Total wighted							
(5*) sum score	571.7	17.3	570.8	19.9	570.7	18.8	
Missing cases 0							

	Intake-course 1977/80						
	Male (n=16)		Female (n=25)		Total (n=41)		2-tail t-test
Variables	x	sd	x	sd	x	sd	p
(1) Stage I: sum scores	52.8	5.8	52.8	6.1	52.8	5.9	
Stage II:							
(2) Theorytest scores	12.3	1.3	12.9	1.3	12.6	1.7	
(3) Practice test sum scores	30.5	2.1	30.5	1.8	30.5	1.9	
(4) Teaching episode *) sum score							
Total wighted	14.3	0.6	14.5	0.6	14.4	0.6	
(5*) sum score	57.1	2.3	58.0	2.4	57.7	2.4	
Missing cases 0							

	Intake-course 1986/88						
	Male (n=21)		Female (n=21)		Total (n=42)		2-tail t-test
Variables	x	sd	x	sd	x	sd	p
(1) Stage I: sum scores	29.9	2.6	32.4	2.5	31.1	2.8	***
Stage II:							
(2) Theorytest scores	36.7	8.3	38.8	5.1	37.7	6.9	
(3) Practice test sum scores	86.5	12.3	86.6	7.9	86.5	10.2	
(4) Teaching episode *) sum score	31.8	6.3	27.7	4.0	29.8	5.6	**
Total wighted							
(5*) sum score	184.9	25.6	185.5	14.0	189.1	20.4	
Missing cases 0							

**=p<0.01
***=p<0.001

(5*) = total score contains additional intake points and music test scores not coded and analyzed in this study.

**=p<0.01
***=p<0.001

APPENDIX 9.3.2 Comparison of the final grades of the course of didactic observation and microteaching of the students' grouped to decade 1970's and 1980's between male, female and total population groups; two-tailed t-tests

Variables	1970-1980 Male			1970-1980 Female			1970-1980 Total		
	n=47	n=37	t	n=75	n=46	t	n=122	n=83	t
Theory	21.0	20.3		22.8	22.8		22.1	21.3	
	(5.0)	(5.1)		(4.7)	(5.1)		(4.8)	(5.2)	
Practice	23.2	21.5		24.4	23.3		23.9	22.5	*
	(4.4)	(4.3)		(3.8)	(4.6)		(4.0)	(4.6)	
Total	22.6	20.4		24.2	22.5	*	23.6	21.8	**
	(4.3)	(4.3)		(3.9)	(4.8)		(4.1)	(4.7)	

* = $p<.05$; ** = $p<.01$

APPENDIX 9.4.1 Pearson's correlation coefficients between selection variables for male and female in four intake-course (n = 205)

Variables	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
1975/76		Men (n=26)					Women (n=43)			
(1) Stage I: sum score	-					-				
Stage II:										
(2) Theorytest score	.85**	-				.78**	-			
(3) Practice test sum score	-.63**	-.57**	-			-.62**	.43**	-		
(4) Teaching episode *) sum score	-.09	-.05	.28	-		-.19	-.15	.22	-	
(5) Sum score	.39*	.34	.42*	.22	-	.64**	.75**	-.04	.00	-
1976/79		Men (n=21)					Women (n=32)			
(1) Stage I: sum score	-					-				
Stage II:										
(2) Theorytest score	.98**	-				.97**	-			
(3) Practice test sum score	-.60**	-.57**	-			-.42*	-.46**	-		
(4) Teaching episode *) sum score	-.55**	-.58**	.02	-		-.33	-.37*	-.16	-	
(5) Sum score	.20	.21	.09	.46*	-	.19	.13	.56**	.28	-
1977/80		Men (n=16)					Women (n=25)			
(1) Stage I: sum score	-					-				
Stage II:										
(2) Theorytest score	.93**	-				.87**	-			
(3) Practice test sum score	-.51*	-.59**	-			-.36	-.17	-		
(4) Teaching episode *) sum score	-.17	-.02	.36	-		-.07	-.16	.34	-	
(5) Sum score	-.02	-.11	.84**	.44	-	.27	.43*	.76**	.22	-
1986/88		Men (n=21)					Women (n=21)			
(1) Stage I: sum score	-					-				
Stage II:										
(2) Theorytest score	-.10	-				-.07	-			
(3) Practice test sum score	.10	.54**	-			.11	.02	-		
(4) Teaching episode *) sum score	-.37	-.24	.21	-		.03	.13	.12	-	
(5) Sum score	-.03	.80**	.90**	.19	-	.01	.69**	.70**	.38	-

* = significant at the 5% level (2-tailed), ** = significant at the 1% level (2-tailed)

*) teaching episode –76 = oral presentation, 7 p.

APPENDIX 9.4.2 Pearson's correlation coefficients between criterion variables: Students marks in the course of microteaching: (1) theory scores, (2) practice scores, (3) final marks and the final mark of the course of didactic observation among course gender group students, and for decade 1970's and 1980's gender course groups, n = 205

Variables	Male			Female			
	1	2	3	1	2	3	All
Course: 1976							
M n=26	.37	.57	.60**	.34	.11	.13	.33*
F n=43							
Course 1979							
M n=21	.26	.12	.35	.16	.22	.09	.21
F n=32							
Course 1980							
M n=16	.53	.11	.18	.86**	.78	.94**	.69***
F n=25							
Course 1988							
M n=21	.22	.23	.58**	.31	.16	.64**	.61**
F n=21							
1970 n=122							
M n=47			.29			.12	.29
F n=75							
1980 n=83							
M n=37			.62***			.90***	.90***
F n=46							
1976 – 1988 n=205			.58**			.47**	.53**
M n=84							
F n=121							

* = significant at the 5% level
** = significant at the 1% level
*** = significant at the 0.1 level

APPENDIX 9.5.1 Pearsons correlations coefficients between predictors and criterion variables: (5) the final mark in the didactic observation and microteaching course in four intake course, n = 205

Variables	1974/76 (N=69)	1976/79 (N=53)	1977/80 (N=41)	1986/88 (N=42)
STUDENTS ENTER CHARACTERISTICS				
SELECTION PROCEDURE:				
S11 stage I: sum score stage II:	.13	-.01	.21	.26
S12 theory test score	.10	.04	.13	.35*
S13 practice test sum score	-.24*	-.05	-.18	.14
S14 teaching episode *) sum score	.13	.14	-.21	-.08
S15 (the total score of the selection procedure)	(.00)	(.13)	(-.04)	(.28)
STUDENTS ATTITUDES EXPECTATIONS CONCERNING "IDEAL" PE-TEACHER CHARACTERISTICS:				
I31 F1 Teacher's congruence/ genuineness (-)	.07	-.07	.00	-.28
I32 F2 Social form: class-centered (+) - individual centered (-)	-.03	-.26	.09	-.14
I33 F3 Teacher involvement (+)	-.01	.14	.19	-.00
I34 F4 Teacher fact centeredness (+) vs. student centeredness (-)	.14	.16	.11	-.08
STUDENTS TEACHING BEHAVIOR (CONTROL): RATED TEACHING EPISODE 5 MIN:				
R41 Item1: presentation	.16	.37**	.06	-.09
R42 Item2: understanding of task content	.14	.36**	.22	-.01
R43 Item3: teacher-pupil interaction	.18	.39**	.06	.00
R44 Item4: creativity	.10	.38**	.31*	.17
R45 (sum score)	(.18)	(.47**)	(.20)	(.04)
STUDENTS PROCESS-BEHAVIOR (CONTROL)				
51 F1: teacher initiation (-) vs. response behavior (+) **)	(.31*)	-	-	-.02
52 ID-index **)	(.18)	-	-	.05

* = significant at the 5% level (2-tailed)
** = significant at the 1% level (2-tailed)
( ) = not used as predictors in multiple regression analysis
*) teaching episode -76 = oral presentation, 7 p.
**) measured -76, not used in these analysis

APPENDIX 9.5.1.1 Pearsons correlations coefficients between predictors and criterion variables: students teaching behavior (mean of microlesson 2 and 3) (1) F-1 score, (2) ID-index, (3) theory test score, (4) the final mark of practice and (5) the final mark in the didactic observation and microteaching course, intake course 1986/1988, male, n = 21

Variables	(1)	(2)	(3)	(4)	(5)
STUDENTS ENTER CHARACTERISTICS					
SELECTION PROCEDURE:					
S11 stage I: sum scores stage II:	-.08	-.09	-.25	-.17	-.23
S12 theory test score	.23	-.04	.42	.26	.38
S13 practice test sum scores	.38	.24	.22	.06	.14
S14 teaching episode sum scores	.42	.35	.13	-.12	-.09
S15 (the total scores of the selection procedure)	(.41)	(.19)	(.41)	(.17)	(.28)
STUDENTS ATTITUDES EXPECTATIONS CONCERNING "IDEAL" PE-TEACHER CHARACTERISTICS:					
I31 F1 Teacher's congruence/ genuineness (-)	-.19	-.26	-.12	-.59**	-.54*
I32 F2 Social form: class-centered (+) - individual centered (-)	-.14	-.12	-.21	-.03	-.10
I33 F3 Teacher involvement (+)	-.11	.02	-.00	.10	.01
I34 F4 Teacher fact centeredness (+) vs. student centeredness (-)	.19	.08	.11	-.04	.02
STUDENTS TEACHING BEHAVIOR (CONTROL): RATED TEACHING EPISODE 5 MIN:					
R41 Item1: presentation	.10	.26	.10	-.03	.10
R42 Item2: understanding of task control	.24	.40	.13	-.06	.14
R43 Item3: teacher-pupil interaction	-.13	.11	.19	.16	.29
R44 Item4: creativity	.14	.26	.28	.16	.34
R45 (sum score)	(.13)	(.30)	(.22)	(.08)	(.27)
STUDENTS PROCESS-BEHAVIOR (CONTROL)					
51 F1: teacher initiation (-) vs. response behavior (+)	.01	.14	.08	.10	.06
52 ID-index	-.10	.12	-.09	.05	.01

* = significant at the 5% level (2-tailed)
** = significant at the 1% level (2-tailed)
( ) = not used as predictors

APPENDIX 9.5.1.2 Pearsons correlations coefficients between predictors and criterion variables: students teaching behavior (mean of microlesson 2 and 3) (1) F-1 score, (2) ID-index, (3) theory test score, (4) the final mark of practice and (5) the final mark in the didactic observation and microteaching course, intake course 1986/1988, female, n = 21

Variables	(1)	(2)	(3)	(4)	(5)
STUDENTS ENTER CHARACTERISTICS					
SELECTION PROCEDURE:					
(a) stage I: sum score stage II:	-.35	-.44*	.58**	.52*	.47*
(b) theory test score	.16	.29	.02	.28	.23
(c) practice test sum score	-.22	-.17	.15	.23	.16
(d) teaching episode sum score	.20	.34	.18	.43	.34
(the total score of the selection procedure)	(.00)	(.17)	(.10)	(.41)	(.30)
STUDENTS ATTITUDES EXPECTATIONS CONCERNING "IDEAL" PE-TEACHER CHARACTERISTICS:					
(e) F1 Teacher's congruence/ genuineness (-)	-.12	-.28	.20	.05	.03
(f) F2 Social form: class-centered (+) - individual centered (-)	.20	.24	-.13	-.05	-.13
(g) F3 Teacher involvement (+)	.41	.43	.20	.06	.19
(h) F4 Teacher fact centeredness (+) vs. student centeredness (-)	-.18	-.22	-.03	-.16	-.11
STUDENTS TEACHING BEHAVIOR (CONTROL): RATED TEACHING EPISODE 5 MIN:					
(i) Item1: presentation	.45*	.54*	-.12	.03	-.01
(j) Item2: understanding of task content	-.08	.16	-.16	.15	.07
(k) Item3: teacher-pupil interaction	.08	.34	-.12	.19	.14
(l) Item4: creativity	.09	.25	-.13	.19	.19
(sum score)	(.15)	(.40)	(-.17)	(.19)	(.14)
STUDENTS PROCESS-BEHAVIOR (CONTROL)					
(m) F1: teacher initiation (-) vs. response behavior (+)	.36	.27	.15	.33	.29
(n) ID-index	.15	-.04	.16	.33	.30

* = significant at the 5% level (2-tailed)
** = significant at the 1% level (2-tailed)
( ) = not used as predictors

APPENDIX 9.6.1.1 Results of regression analyses for the male students intake course 74/76 (n = 26). Regression coefficient (b), standard errors in brackets and standardized regression coefficients (β)

Predictor variables	**Criterion variables**					
	(Theory scores)		Practice scores		Final mark	
	b	β	b	β	b	β
	x)					
Selection variable						
Stage II						
- theory test scores			.05(.03)	.31		
Stage II						
- teaching episode scores					.67(2.8)	.42*
Students' entry rated teaching behavior (control)						
- item 3, teacher-pupil interaction			2.83(.94)	.49**		
Students' attitudes expectations concerning characteristics of "Ideal" P.E. teacher: Factor IV scores "Teacher fact-centeredness (+) vs. Student-centeredness (-)					2.88(1.22)	.42**
Constant			11.50(3.39)		12.31(4.48)	
R	-		.64		.54	
R^2	-		.42		.30	
F	-		8.17**		4.86*	
Classification power	-		69%**		77%**	

* = p<.05
** = p<.01
x) = regression model not selected

APPENDIX 9.6.1.2 Results of regression analyses for the female students intake course 74/76 (n = 43). Regression coefficient (b), standard errors in brackets and standardized regression coefficients (β)

Predictor variables	**Criterion variables**					
	(Theory scores)		Practice scores		Final mark	
	b	β	b	β	b	β
	x)				x)	
Students' entry rated teaching behavior (control)						
- item 1, clarify of presentation			2.77(.68)	0.52***		
Students' attitudes expectations concerning characteristics of "Ideal" P.E. teacher: Factor II, social form: Class-centered (+) vs. Individual-centered (-)			-1.69(.71)	0.31*		
Factor III, teacher involvement (+)			-1.77(.85)	-0.27*		
Constant			13.46(2.54)			
R	-		0.61		-	
R^2	-		0.37		-	
F	-		7.53***		-	
Classification power	-		65%*		-	

* = p<.05
*** = p<.001
x) = regression model not selected

APPENDIX 9.6.2.1 Results of regression analyses for the male students intake course 76/79 (n = 21). Regression coefficient (b), standard errors in brackets and standardized regression coefficients (β)

Predictor variables	**Criterion variables**					
	(Theory scores)		Practice scores		Final mark	
	b	β	b	β	b	β
	x)					
Selection variable						
Stage II						
- practice test scores			-.08(.04)	-.41*		
Students' attitudes						
expectations concerning characteristics of "Ideal" P.E. teacher: Factor I scores "Teachers' congruence/ genuineness (-)"					-2.58(.91)	-.58**
- F-2, Social form: "Class-centered (+) - Individual centered (-)"					-1.60(.69)	-.48*
- F-4, "Teacher fact-centeredness (+) - Students-centeredness (-)"			1.99(.97)	.40*		
Constant			48.40(11.86)		21.60(.66)	
R	-		0.55		0.60	
R^2	-		0.30		0.36	
F	-		3.90*		4.99*	
Classification power	-		67%		71%	

* = p<.05
x) = regression model not selected

APPENDIX 9.6.2.2 Results of regression analyses for the female students intake course 76/79 (n = 32). Regression coefficient (b), standard errors in brackets and standardized regression coefficients (β)

Predictor variables	**Criterion variables**					
	Theory scores		Practice scores		Final mark	
	b	β	b	β	b	β
Students' entry rated teaching behavior (control)						
- item 2, understanding of task content			1.29(.56)	0.39*		
- item 4, creativity	2.31(.47)	0.65***			1.63(.41)	0.58***
Students' attitudes						
expectations concerning characteristics of "Ideal" P.E. teacher: Factor I scores						
- F-2, Social form: "Class-centered (+) - Individual centered (-)"	-1.73(.82)	-0.28*			-1.27(.71)	-0.26
Constant	14.78(1.70)		20.09(2.46)		18.75(1.48)	
R	0.70		0.39		0.63	
R^2	0.50		0.15		0.40	
F	14.32***		5.33*		9.50***	
Classification power	84%***		63%		78%**	

* = p<.05
** = p<.01
*** = p<.001

APPENDIX 9.6.3.1 Results of regression analyses for the male students intake course 77/80 (n = 16). Regression coefficient (b), standard errors in brackets and standardized regression coefficients (β)

Predictor variables	**Criterion variables**					
	Theory scores		Practice scores		Final mark	
	b	β	b	β	b	β
	x)					
Selection variable						
Stage I						
- school success			-.41(.15)	-0.58*		
Students' entry rated teaching behavior (control)						
- item 4, creativity					1.17(.52)	.51*
Students' attitudes expectations concerning characteristics of "Ideal" P.E. teacher: Factor II scores: Social form: "Class-centered (+) - Individual centered (-)"			2.57(1.42)	0.38		
Constant			45.50(7.94)		19.22(1.70)	
R	-		0.65		0.51	
R^2	-		0.43		0.26	
F	-		4.85*		5.02*	
Classification power	-		81%**		62%*	

* = p<.05
** = p<.01
*** = p<.001
x) = regression model not selected

APPENDIX 9.6.3.2 Results of regression analyses for the female students intake course 77/80 (n = 25). Regression coefficient (b), standard errors in brackets and standardized regression coefficients (β)

Predictor variables	**Criterion variables**					
	(Theory scores)		Practice scores		Final mark	
	b	β	b	β	b	β
			x)			
Selection variable						
Stage I						
- school success					.29(.16)	.35
Stage II						
- practice test scores	-.96(.44)	-.37*				
Students' entry rated teaching behavior (control)						
- item 2, understanding of task content	2.15(.87)	.43**				
Students' attitudes expectations concerning characteristics of "Ideal" P.E. teacher: Factor IV scores "Teacher fact-centeredness (+) vs. Student-centeredness (-)	2.41(1.28)	.33				
Constant	57.83(1.28)				6.89(8.53)	
R	.61	-			.35	
R^2	.38			-	.12	
F	4.23*			-	3.26	
Classification power	72%			-	68%	

* = p<.05
x) = regression model not selected

APPENDIX 9.6.4.1 Results of regression analyses for the male students intake course 86/88 (n = 21). Regression coefficient (b), standard errors in brackets and standardized regression coefficients (β)

Predictor variables	**Criterion variables**					
	Theory scores		Practice scores		Final mark	
	b	β	b	β	b	β
Selection variable						
Stage II - theory test scores	.28(.14)	0.42*				
Students' rated entry teaching behavior (control) - item 4, creativity					1.40(.62)	0.40*
Students' attitudes expectations concerning characteristics of "Ideal" P.E. teacher: Factor I scores: "Teachers' congruence/ genuineness (-)"			-3.21(1.02)	-0.59**	3.46(1.10)	-0.58**
Constant	9.15(5.25)		21.24(.83)		15.40(2.39)	
R	0.42		0.59		0.67	
R^2	0.18		0.34		0.45	
F	4.13*		9.98**		7.35**	
Classification power	67%*		62%		71%	

* = p<.05
** = p<.01

APPENDIX 9.6.4.2 Results of regression analyses for the female students intake course 86/88 (n = 21). Regression coefficient (b), standard errors in brackets and standardized regression coefficients (β)

Predictor variables	**Criterion variables**					
	Theory scores		Practice scores		Final mark	
	b	β	b	β	b	β
Selection variable						
Stage I - school success 1.06(0.34)	0.58**	1.58(0.33)	0.90***	1.07(.37)	.57**	
Stage II - theory test scores			.25(0.13)	0.30		
Students' rated entry teaching behavior (control) - item 1, clarity of presentation			2.92(1.20)	0.46*		
- item 4, creativity			1.66(0.81)	0.33*	1.90(1.09)	.35
Constant	-11.51(11.17)		-54.29(15.21)		-17.58(13.51)	
R	.58		.80		.58	
R^2	.33		.63		.34	
F	9.46**		6.88**		4.56*	
Classification power	71%*		81%**		76%*	

* = p<.05
** = p<.01
*** = p<.001

APPENDIX 9.6.4.3 Results of regression analyses for the male students intake course 86/88. Regression coefficient (b), standard errors in brackets and standardized regression coefficients (β)

Predictor variables	**Criterion variables**					
	Measured teaching behavior: ("teacher initiation (-) vs. response bahavior (+)")				Theory &practis scores	
	F-1 scores		Corrected ID-index		Final Mark	
	b	β	b	β	b	β
Selection variable						
(2) Stage II - theory test scores	0.2 (.01)	34				
Teaching episode total scores	.07 (.02)	.69**				
Students' entry rated teaching behavior (control 5 min)						
- item 1 (clarity of presentation)			3.73 (1.96)	.36*		
- item 2 (understanding of task content)			15.22 (3.41)	1.20***		
- item 3 (teacher-pupil interaction)			-15,89 (3.99)	-1.20***		
- item 4 (creativity)					1.40 (.62)	.40*
(6) Students' measured teaching behavior, ID-index			.49 (.16)	.64**		
(7) Students' Attitudes expectations concerning characteristics of "ideal" P.E. teacher: Factor I scores "Teacher's congruence/ genuineness (-)"	-.35 (.10)	-44*	-7.50 (2.08	-.60**	-3.64 (1.10)	-.58**
Constant	-2.33 (.86)		20.26 (10.5)		15.40 (2.39)	
R	.67		.81		.67	
R^2	.45		.66		.45	
F	4.63**		5.83**		7.35**	
Classification power	66,67 %*		85,71 %**		71,43*	

* = p< .05
** = p< .01
*** = p< .001

APPENDIX 9.6.4.4 Results of regression analyses for the female students intake course 86/88. Regression coefficient (b), standard errors in brackets and standardized regression coefficients (β)

Predictor variables	Criterion variables: Measured teaching behavior: ("teacher initiation (-) vs. response behavior (+)") F-1 scores		Corrected ID-index		Theory & practice scores Final Mark	
	b	β	b	β	b	β
Selection variable						
(1) Stage I School success					1.07 (.37)	.57**
Students' entry rated teaching behavior: (control 5 min.)						
- item 1 (clarity of presentation)	48 (.21)	.45*	9.86 (3.49)	.54**		
- item 4 (creativity)					1.90 (1.09)	.35
Constant	-1.54 (.83)		11.82 (13.50)		-17.58 (13.51)	
R	.45		.54		.58	
R^2	.21		.30		.34	
F	4.94*		7.97**		4.56*	
Classification Power	61,90 %		80,95 %**		76,19%*	

* = p< .05
**= p< .01

APPENDIX 10 **Student program evaluation:**
APPENDIX 10.1 Questionnaire (see appendix 5.1)
APPENDIX 10.2 Reliability of students' course rating questionnaire: Cronbach's Alpha of seven varimax factor's sum scores across the four microteaching course group, n = 197

Factor no	% of total variance	number of items	Alpha
1.	24.6	21	.92
2.	6.2	13	.68
3.	3.7	10	.84
4.	2.1	7	.79
5.	2.3	2	.70
6.	2.1	3	.70
7.	1.8	2	.50

% of common variance 43.8

APPENDIX 10.3 Comparison of curriculum groups 1. before study reform (76, 79, 80) and 2. after study reform (81, 82, 88) students' rating of the microteaching course: means, standard deviations and t-tests for the male, female and total populations

Item No.	Male (1) x̄ sd	Male (2) x̄ sd	Total Df= 111	p	Female (1) x̄ sd	Female (2) x̄ sd	Total Df= 168	p	Total (n=286) (1) x̄ sd	Total (n=286) (2) x̄ sd	Total Df= 281	p
2. The course was pretended so that I was aware of its contents and extent from the beginning	2.9 1.1	2.9 1.1	2.9 1.1		2.7 1.3	2.7 1.1	2.7 1.2		2.8 1.0	2.8 1.0	2.8 1.2	
3. I was able to get the right idea of the objectives of the lecture course from the beginning	2.7 1.1	2.7 1.1	2.7 1.0		2.4 1.2	2.4 1.0	2.4 1.1		2.5 1.2	2.5 1.0	2.5 1.1	
4. I was aware of the objectives of the exercises from the beginning	3.4 1.2	3.4 1.2	3.4 1.3		3.0 1.3	3.1 1.2	3.0 1.2		3.2 1.3	3.2 1.2	3.2 1.3	
5. The main concepts of the course were badly presented	2.9 1.0	2.6 0.9	2.8 1.0		2.7 1.2	2.9 1.1	2.8 1.1		2.8 1.1	2.7 1.0	2.8 1.1	
6. The course has awoken an interest in me in this subject	3.4 1.2	2.9 1.2	3.2 1.2	*	3.6 1.0	3.3 1.2	3.5 1.1		3.5 1.1	3.1 1.2	3.4 1.2	**
7. I did not learn much during the lectures	3.0 1.2	2.6 1.1	2.8 1.2		2.7 1.1	3.2 1.0	2.9 1.1	**	2.8 1.1	3.0 1.1	2.9 1.1	
8. I did not learn much during the exercises	2.0 1.2	2.3 1.2	2.1 1.2		2.0 1.0	2.1 1.1	2.0 1.0		2.0 1.1	2.2 1.2	2.1 1.1	
9. Exercise tasks have been sensible	3.3 1.1	3.1 1.0	3.2 1.1		3.5 1.1	3.3 1.1	3.4 1.1		3.4 1.1	3.2 1.1	3.3 1.1	
10. Using students as "pupils" has been reasonable	2.9 1.5	3.7 1.4	3.2 1.5	**	3.4 1.5	3.4 1.3	3.4 1.4		3.2 1.5	3.5 1.4	3.3 1.4	
11. The course has overlapped unnecessarily with my earlier studies	2.0 0.9	2.5 1.1	2.2 1.0	**	1.7 0.8	2.2 1.1	1.9 0.9	***	1.8 0.8	2.3 1.1	2.0 1.0	***
12. The course was sensible linked with earlier studies	3.4 1.1	3.3 1.1	3.3 1.1		3.5 1.0	3.3 1.0	3.4 1.0		3.4 1.1	3.3 1.0	3.4 1.1	
13. The course was organized well compared with other corresponding cource	2.8 1.0	2.9 0.8	2.9 0.9		3.0 1.0	2.8 0.8	2.9 0.9		2.9 1.0	2.9 0.8	2.9 0.9	
14. The contents of the lecture and the exercises did not match with each other sufficiently	2.6 1.1	2.6 1.2	2.6 1.1		2.6 1.2	3.0 1.1	2.7 1.2		2.6 1.2	2.8 1.1	2.7 1.2	
15. This course should have been placed earlier in the study programme	2.4 1.4	2.4 1.2	2.4 1.3		2.3 1.4	3.0 1.3	2.6 1.4	**	2.4 1.4	2.7 1.3	2.5 1.4	*
16. Exercises contained too few tasks of different types	2.8 1.4	2.5 1.3	2.7 1.3		2.5 1.3	2.7 1.2	2.6 1.2		2.6 1.3	2.6 1.2	2.6 1.3	
17. Exercises proceeded too quickly	2.5 1.3	2.9 1.4	2.7 1.4		2.2 1.3	2.9 1.3	2.5 1.3	***	2.3 1.3	2.9 1.4	2.6 1.3	***
18. Too little time was spent on the analysis of feedback	3.0 1.5	2.6 1.1	2.8 1.3		2.7 1.4	2.8 1.3	2.7 1.3		2.8 1.4	2.7 1.2	2.8 1.3	
19. Lectures should have included more audiovisual equipment	3.1 1.3	3.2 1.2	3.1 1.3		2.5 1.2	2.8 1.2	2.6 1.2		2.7 1.3	2.9 1.2	2.8 1.3	

table continues

Item No.	Male (1) x̄ sd	Male (2) x̄ sd	Total Df= 111	p	Female (1) x̄ sd	Female (2) x̄ sd	Total Df= 168	p	Total (n=286) (1) x̄ sd	Total (n=286) (2) x̄ sd	Total Df= 281	p
20. Lecturer has been to detach (impersonal)	2.0 1.0	1.9 1.0	2.0 1.0		1.7 0.8	2.0 0.9	1.8 0.9	*	1.8 0.9	1.9 1.0	1.9 0.9	
21. Lecturer has spoken loud enough	3.1 1.3	2.8 1.1	2.9 1.3		2.5 1.3	2.6 1.2	2.5 1.2		2.7 1.3	2.6 1.2	2.7 1.2	
22. Throughout the semester I remained unware of the objectives of the course	2.8 1.3	2.5 1.2	2.7 1.3		2.2 1.1	2.6 1.1	2.3 1.1		2.4 1.2	2.6 1.1	2.5 1.2	
23. The main concepts of the course have been presented clearly enough	3.3 1.1	3.3 1.1	3.3 1.1		3.4 1.1	3.1 1.1	3.3 1.1		3.4 1.1	3.2 1.1	3.3 1.1	
24. Lecturer did not give the students enough time to ask questions	2.5 1.1	2.4 1.0	2.4 1.1		2.5 1.1	2.5 1.1	2.5 1.1		2.5 1.1	2.5 1.1	2.5 1.1	
25. During the course teachers were careless as regards deadlines for assignments	2.1 1.1	2.2 1.0	2.1 1.0		1.8 1.1	2.2 1.1	2.0 1.1		1.9 1.1	2.2 1.1	2.0 1.1	*
26. I was generally bored during lectures	3.8 1.9	3.1 1.0	3.5 1.0	***	2.9 1.1	3.4 1.0	3.1 1.1	***	3.2 1.1	3.3 1.0	3.3 1.1	
27. I was generally bored during exercises	2.4 1.3	2.4 1.1	2.4 1.2		1.9 1.1	2.6 1.3	2.1 1.2	***	2.1 1.2	2.5 1.2	2.2 1.2	**
28. Handouts outlining the contents of lectures were useful from the point of view of attaining the objectives of lectures	4.3 1.1	3.8 1.1	4.1 1.1	*	4.6 0.7	4.1 1.0	4.4 0.9	***	4.5 0.9	4.0 1.1	4.3 1.0	***
29. It was difficult to follow the lecture	3.5 1.2	3.1 1.1	3.4 1.1	*	3.5 1.2	3.6 1.2	3.5 1.2		3.5 1.2	3.4 1.2	3.4 1.2	
30. The whole course is useless in educating P.E. teachers	2.0 1.2	2.1 1.2	2.0 1.2		1.7 1.9	1.8 1.0	1.8 0.9		1.8 1.0	1.9 1.1	1.9 1.0	
31. Lecturer should have proceeded more quickly	2.0 0.9	2.3 1.1	2.1 1.0	*	1.9 1.1	2.5 1.0	2.1 1.0	***	1.9 0.9	2.4 1.0	2.1 1.0	***
32. Lecturer did not know the subjects well enough	1.9 1.1	1.7 0.9	1.8 1.0		1.6 0.9	1.6 0.8	1.6 0.8		1.7 1.0	1.6 1.8	1.7 0.9	
33. Lecture's personal opinions biased teaching too much	3.0 1.2	3.0 1.2	3.0 1.2		2.8 1.2	3.1 1.0	2.9 1.1		2.9 1.2	3.1 1.1	3.0 1.2	
34. Time reserved for exercises was usually too short	3.2 1.4	3.0 1.3	3.1 1.4		3.1 1.5	3.5 1.3	3.3 1.4		3.2 1.5	3.3 1.3	3.2 1.4	
35. The course as such is rather useful	4.0 1.0	3.7 1.2	3.9 1.1		4.1 0.9	3.9 1.0	4.0 1.0		4.1 1.0	3.8 1.1	4.0 1.0	*
36. The course did not deal with really essential and important matters	3.0 1.3	2.7 1.2	2.9 1.3		2.6 1.3	2.6 1.1	2.6 1.2		2.8 1.3	2.6 1.2	2.7 1.3	
37. Lectures and exercises were integrated well	3.2 1.1	3.5 1.1	3.3 1.1		3.3 1.1	3.2 1.1	3.3 1.1		3.3 1.1	3.3 1.1	3.3 1.1	
38. It was easy to keep interested in the subjects during the lectures	2.1 1.1	2.7 1.1	2.4 1.1	**	2.3 1.0	2.4 1.0	2.4 1.0		2.3 1.0	2.5 1.0	2.4 1.0	*
39. It was easy to keep interested in the subjects during the exercises	3.5 1.1	3.3 1.3	3.4 1.2		4.0 1.0	3.3 1.2	3.7 1.1	***	3.8 1.1	3.3 1.2	3.6 1.2	**
40. The course did not awake any interest in me in the subject	2.7 1.2	2.9 1.3	2.8 1.2		2.2 1.0	2.7 1.2	2.4 1.1	**	2.4 1.1	2.8 1.2	2.5 0.2	**

table continues

Item No.	Male (1) x̄ sd	Male (2) x̄ sd	Total Df= 111	p	Female (1) x̄ sd	Female (2) x̄ sd	Total Df= 168	p	Total (n=286) (1) x̄ sd	Total (n=286) (2) x̄ sd	Total Df= 281	(* p
41. Lecturer has pointless habits and mannerism which divert the student's attention from teaching	3.1 1.2	2.8 1.3	3.0 1.2		2.8 1.2	3.3 1.2	3.0 1.2		2.9 1.2	3.1 1.2	3.0 1.2	
42. This course should have been placed later in the study programme	2.1 1.1	2.1 0.9	2.1 1.0		1.8 1.0	1.6 0.8	1.7 0.9		2.0 1.1	1.8 0.9	1.9 1.0	-
43. I have learnt more in the lectures of this course than in lectures in general	2.1 0.9	2.5 0.8	2.3 0.9	*	2.1 0.9	2.3 1.0	2.2 1.0		2.1 0.9	2.4 0.9	2.2 0.9	
44. Lecturer did not take the students into consideration well enough	3.2 1.1	2.8 1.1	3.0 1.1		2.8 1.1	2.7 1.0	2.8 1.1		3.0 1.1	2.8 1.0	2.9 1.1	
45. Handouts summarizing the main points of lectures were useless	1.5 0.8	2.1 1.1	1.7 1.0	***	1.4 0.7	1.8 1.0	1.6 0.8	***	1.4 0.7	1.9 1.0	1.6 0.9	***
46. Lecture course gave me new ideas about P.E. teaching	3.3 1.2	3.6 1.0	3.4 1.1		3.8 1.1	3.4 1.1	3.6 1.1		3.6 1.1	3.5 1.0	3.5 1.1	
47. Demonstration tasks were badly selected	2.4 1.0	2.5 1.1	2.4 1.1		2.2 0.9	2.1 0.9	2.1 0.9		2.3 1.0	2.2 1.0	2.3 1.0	
48. Lecture course was not worth attending	2.5 1.1	2.5 1.1	2.5 1.1		2.2 0.9	2.8 1.0	2.4 1.0	***	2.3 1.0	2.6 1.1	2.4 1.0	**
49. From the point of view of educating P.E. teachers it would have been more useful to spend the time on other types of teaching practise	2.5 1.2	2.5 1.2	2.5 1.2		2.3 1.3	2.5 1.2	2.4 1.2		2.4 1.2	2.5 1.2	2.4 1.2	
50. Demonstrations of lecture and teaching models would be sufficient without having to participate in exercises	1.7 1.0	1.8 1.0	1.8 1.0		1.5 0.9	1.7 0.9	1.6 0.9		1.6 1.0	1.7 1.0	1.7 1.0	
51. Lecturer proceeded too quickly	2.8 1.1	2.8 1.1	2.8 1.1		2.7 1.1	2.6 1.1	2.6 1.1		2.7 1.1	2.7 1.0	2.7 1.1	
52. Organization of exercises was not good enough	2.3 1.1	2.7 1.1	2.5 1.1		2.3 1.1	2.3 1.1	2.3 1.1		2.3 1.1	2.4 1.1	2.4 1.1	
53. The teaching skills of the exercise supervisor were not good enough	2.0 1.0	2.0 1.1	2.0 1.0		1.9 0.9	2.0 0.9	1.9 0.9		1.9 1.0	2.0 1.0	2.0 1.0	
54. The exercise tasks were explained clearly	3.4 1.0	3.3 1.3	3.4 1.2		3.3 1.2	3.0 1.3	3.2 1.2		3.4 1.2	3.1 1.3	3.3 1.2	
55. I learned to distinguishes teaching models observing and classifying feedback	3.8 1.1	4.0 1.0	3.9 1.0		4.0 1.2	4.0 1.1	4.0 1.1		3.9 1.0	4.0 1.0	3.9 1.0	
56. Exercises clarified the issues presented in lectures	3.7 1.0	4.0 0.8	3.8 0.9		4.1 0.9	3.8 1.0	4.0 0.9		4.0 1.0	3.9 0.9	3.9 1.0	
57. I believe I have obtained a broader view of teaching behaviour	4.0 1.0	4.0 0.9	4.0 1.0		4.4 1.2	4.0 1.3	4.2 1.2	**	4.2 0.9	4.0 1.0	4.1 0.9	*
58. I will probably use the various teaching models presented consciously in my teaching	3.7 1.1	3.9 1.2	3.7 1.1		4.2 0.7	3.8 1.1	4.0 0.9	**	4.0 0.9	3.8 1.1	3.9 1.0	
59. I became aware of my personal teaching defects inadequates during the course	3.9 1.0	3.6 1.2	3.7 1.1		3.7 1.2	4.0 1.0	3.8 1.1		3.8 1.1	3.8 1.1	3.8 1.1	

*, **, ***, $p<0.05$, 0.01 and 0.001 respectively

(* = 3 cases were excluded from the analysis because they had at least one missing discriminant variable

APPENDIX 10.4 Standardized canonical discriminant function coefficients and univariate F-ratios based on students' course ratings, students grouped to two course groups 1. before and 2. after study reform (n = 283)

Variables item	Male (n=113)			Female (n=170)			Total (n=283)	
no	Function	F-ratio, df=111	no	Function	F-ratio	no	Function	F-ratio
26	-.58	12.51***	28	.47	19.69***	45	.21	24.11***
45	.64	12.33***	11	-.24	14.21***	11	-.40	21.5***
6	-1.09	4.60*	47	.64	0.55	31	-.32	15.96***
10	.39	7.58**	17	-.46	10.57***	36	.35	0.65
46	.92	12.33***	31	-.27	12.67***	17	-.27	11.92***
52	.36	2.60	57	.29	8.71**	10	-.23	2.94
11	.52	7.27**	42	.38	2.13	37	-.31	0.21
21	-.76	2.19	59	-.32	3.34	47	.31	0.10
22	-.49	1.06	7	-.51	11.08***	55	-.21	0.29
2	-.66	0.13	2	-.43	0.20	23	.24	2.20
41	-.21	0.78	41	-.34	7.08**	44	.23	2.04
28	-.50	4.95*	36	.28	0.13	54	.22	3.31
34	-.39	0.37	13	.39	1.14	28	.36	21.83***
47	.39	0.36	35	.37	3.37	14	-.25	2.11
25	-.23	0.81	43	-.37	1.65	32	.15	0.24
20	.36	0.12	30	.51	0.12	38	-.18	3.72*
55	.58	0.87	44	.27	0.16	42	.21	0.45
37	.59	2.52	48	-.39	12.65***	35	.30	5.69
38	.23	6.96**	39	.33	12.00***	24	.14	0.12
35	-.60	2.18	14	-.18	10.30**	4	-.17	0.44
4	.43	0.46	29	.33	0.35	13	.24	0.49
19	.41	0.42	8	.28	0.43	43	-.17	4.89*
30	-.42	0.30	53	-.22	1.50	41	-.17	1.20
50	.31	0.14	20	.23	3.22*	2	-.16	0.51
42	.30	0.15	50	.22	0.92	6	.24	7.86**
14	.19	0.22	22	-.21	4.64*	39	.21	8.81**
57	.44	0.22	38	-.26	0.69	15	-.12	5.76*
51	-.22	0.84	54	.20	3.26			
40	.24	3.44	45	.24	11.43***			
44	-.19	1.60	16	-.18	1.21			
36	-.26	0.49	52	.17	0.98			
54	-.44	2.79	23	.17	3.46			
			18	-.14	0.25			
Eigenvalue 3.06			Eigenvalue 1.40			Eigenvalue 0.65		
RC .87			RC .76			RC .61		
Wilks' Lambda 0.247			Wilks' Lambda 0.4166			Wilks' Lambda 0.623		
Chi Square 132.32, df=33, sig. 0.000			Chi Square 132.65, df=33, sig. 0.000			Chi Square 126.50, df=27, sig. 0.000		
*p<.05, **p<0.01, ***p<0.001			*p<.05, **p<0.01, ***p<0.001			*p<.05, **p<0.01, ***p<0.001		

APPENDIX 10.5 Factor analysis of students' ratings of the microteaching course (1976, 1979, 1982, 1988), n = 203

Items No	Factor loadings 1	2	3	4	5	6	7	h^2
F1: Course in curriculum program (+) – (-)								
49. From the point of view of educating P.E. teachers it would have been more useful to spend the time on other types of teaching	-.68	-.34	.10	-.17	-.03	.02	.04	.62
30. The whole course is useless in educating P.E. teachers	-.67	-.22	.04	-.13	-.09	.19	.05	.56
35. The course as such is rather useful	.66	.30	.07	.17	.09	.01	.11	.58
9. Exercise tasks have been sensible	.63	.17	.08	.04	-.19	-.03	-.10	.49
40. The course did not awake any interest in me in the subject	-.61	.06	.17	.15	.05	-.20	.43	.65
27. I was generally bored during exercises	-.60	.09	.04	-.31	.18	-.09	.11	.52
39. It was easy to keep interested in the subjects during the exercises presented in lectures	.57	.04	.12	.28	-.19	.07	.04	.46
57. I believe I have obtained a broader view of teaching behaviour	.57	-.01	-.24	.08	.08	-.22	.08	.45
8. I did not learn much during the exercises	-.55	-.04	.11	.24	-.01	.02	.06	.25
50. Demonstrations of lecture and teaching models would be sufficient without having to participate in exercises	-.54	.09	.03	-.08	.03	.04	.08	.31
36. The course did not deal with really essential and important matters	-.53	-.30	.24	-.17	-.18	-.04	.14	.58
46. Lecture course gave me new ideas about P.E. teaching	.50	.10	-.24	.07	.10	-.25	-.29	.48
47. Demonstration tasks were badly selected	-.50	-.12	.21	-.00	.31	.02	.08	.41
6. The course has awoken an interest in me in this subject	.46	.03	-.11	.25	.15	.22	-.35	.49
58. I will probably use the various teaching models presented concsciously in my teaching	.43	.19	-.17	.17	.01	.15	-.36	.43
53. The teaching skills of the exercise supervisor were not good enough	-.39	-.20	.21	-.13	.18	.00	.14	.31
59. I became aware of my personal teaching defects inadequates during the course	.34	.20	-.10	.20	-.01	.05	.08	.21
62. The actual teaching of the planned teaching episode, when only the goal was given, was interested	.33	.28	-.03	.23	-.21	.07	.01	.28
F2: Clarity of goal presentation, (+) – (-)								
3. I was able to get the right idea of the objectives of the lecture course from the beginning	.02	.68	-.08	.04	-.11	.01	-.03	.48
2. The course was pretended so that I was aware of its contents and extent from the beginning	.08	.63	-.17	.03	-.14	.12	.01	.47
4. I was aware of the objectives of the exercises from the beginning	.34	.60	.07	.02	-.16	.03	.04	.44
23. The main concepts of the course have been presented clearly enough	.15	.53	-.18	.19	.06	.22	-.07	.46
38. It was easy to keep interested in the subjects during the lectures	-.10	.46	-.49	.00	.01	-.18	-.20	.53
21. Lecturer has spoken loud enough	.09	.45	-.23	.15	.03	.02	.10	.30
12. The course was sensible linked with earlier studies	.21	.41	-.15	.02	.11	.21	.07	.33
22. Throughout the semester I remained unware of the objectives of the course	-.20	-.39	-.28	-.06	.12	-.11	-.32	.39
54. The exercise tasks were explained clearly	.27	.33	.07	.04	-.32	.04	.01	.30
13. The course was organized well compared with other corresponding course	.25	.31	-.26	.18	.08	.14	.00	.29
41. Lecturer has pointless habits and mannerism which divert the student's attention from teaching	-.14	.26	-.20	-.20	.12	-.11	.22	.23

APPENDIX 10.5 continued

Items No	Factor loadings 1	2	3	4	5	6	7	h^2
F3: Theory and practice integration (+) – (-)								
14. The contents of the lecture and the exercises did not match with each other sufficiently	0.6	-.09	.54	-.01	.05	-.07	.02	.31
29. It was difficult to follow the lecture	.10	-.38	.52	-.14	.10	.15	.08	.48
56. Exercises clarified the issues	.36	-.07	.51	.18	-.05	.28	.09	.51
37. Lectures and exercises were integrated well	.09	.18	-.50	.15	.10	.31	-.06	.43
38. It was easy to keep interested in the subjects during the lectures	.10	.46	-.49	.06	.01	-.18	-.20	.42
48. Lecture course was not worth attending	-.33	-.29	.38	-.06	-.02	-.18	.21	.37
7. I did not learn much during the lectures	-.35	-.22	.38	-.15	-.10	-.02	.12	.37
33. Lecture's personal opinions biased teaching too much	-.24	-.15	.35	-.19	.06	-.05	.19	.28
5. The main concepts of the course were badly presented	-.24	-.27	.29	-.04	.09	-.13	-.06	.25
F4: Structural outine for teaching episodes and feedback (+) – (-)								
63. Filling in the structural outline for a teaching episode was useless	-.22	-.04	.04	-.68	.02	.17	.18	.58
65. The structural outline facilitated the construction of the plan for the teaching episode	.17	.05	-.12	.66	.07	.04	.02	.49
61. The task of evaluating teaching was useful	.26	.16	.10	.47	.13	.03	-.04	.35
55. I learned to distinguishes teaching models	.25	.05	-.11	.39	-.02	.06	-.08	.24
64. The way the course groups were set up was sensible	.22	.16	-.10	.35	-.11	-.07	.04	.25
F5: Time reservation for events (+) – (-)								
17. Exercises proceeded too quickly	.13	.04	.04	.04	.56	-.02	.05	.34
18. Too little time was spent on the analysis of feedback	.01	.01	.05	.08	.53	.01	-.05	.29
34. Time reserved for exercises was usually too short	.03	-.01	-.02	.02	.51	.07	.06	.27
52. Organization of exercises was not good enough	-.09	-.20	.06	.11	-.40	-.12	.14	.26
F6: Handouts lectures (+) – (-)								
28. Handouts outlining the contents of lectures were useful from the point of view of attaining the objectives of lectures	.08	.02	-.06	.06	-.14	.78	-.11	.66
45. Handouts summarizing the main points of lectures were useless	.12	-.16	-.10	-.10	.15	-.61	.04	.48
F7: Use of AV material (+) – (-) in theory practice integrations								
26. I was generally bored during lectures	-.23	-.25	.33	.05	.07	.07	.58	.58
19. Lectures should have included more audiovisual equipment	.04	.01	.07	.05	.33	.12	.42	.31
Eigenvalue	11.6	2.5	1.9	1.7	1.2	1.0	1.0	20.25
~~% common~~ variance	~~55.6~~	12.0	9.0	8.3	~~5.6~~	4.9	4.3	100
% total variance	22.8	4.9	3.5	3.4	2.3	2.0	1.9	41.0

APPENDIX 10.5.1 Students' course ratings factor's transformation matrix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
F1.	.69	.44	-.36	.32	-.09	.19	-.22
F2.	-.55	.64	-.41	-.22	-.10	-.06	-.05
F3.	.02	-.10	-.11	.14	.91	-.35	-.01
F4.	.29	.39	.45	-.14	-.22	-.64	.28
F5.	-.25	-.08	-.20	-.77	-.19	-.90	.46
F6.	-.06	.40	.48	.11	.24	.61	.40
F7.	.27	-.15	-.47	-.43	.06	.08	.70