UNIVERSITY OF JYVÄSKYLÄ
Centre for Applied Language Studies

Sari Luoma

WHAT DOES YOUR TEST MEASURE?

Construct definition in language test development and validation

ISBN 951-39-0897-6 (electronic version)

ABSTRACT

Luoma, Sari
What does your test measure?
Construct definition in language test development and validation.
Jyväskylä: Centre for Applied Language Studies, University of Jyväskylä.
Manuscript.

This study concerns language testing methodology. It focuses on the principles and practices of test development and validation, especially the role of the theoretical construct definition in these processes. The aim of the thesis is to clarify the principles by which test developers can and should build quality into their tests and the practical activities that these principles entail.

The thesis builds on the notion that the construct definition is the most important concern in test development and validation. In language testing, the construct definition has two sides, theoretical and psychometric. The thesis focuses on the role of the theoretical construct definition because while the importance of the theoretical definition is recognised in theoretical texts, the practical activities that it entails for test development are less clearly defined. Since the two sides of the construct definition are entwined, the interface between them is also investigated.

The topic is addressed from two perspectives, recommendations from theory and reports of practice. Theoretical texts on test development, validation, and construct definition are treated with a view to define goals for accountable measurement and recommendations for how they should be addressed. A summary model of test development and validation is developed and used as a guiding framework in an analysis of three cases of reported practice in test development. The cases provide a range of examples of what is considered acceptable practice by different test development boards, and thereby a range of concrete examples for how the goals of desirable measurement can be addressed.

The results of the study highlight the importance of the combination of theoretical and psychometric perspectives in the definition of constructs. This enables test developers to say what the test scores mean as well as prove the measurement quality of the test. It may also enable further development of theories of second/foreign language ability. The combination of theoretical and psychometric construct definitions is an interesting area of future research, particularly in performance-based assessment where the task context is less clearly defined than in traditional measurement. Both the nature of spoken interaction and the assessment challenges that it entails deserve further study. Another important avenue for future research is the provision of ethnographic perspectives into actual practices of examination development.

Keywords: test development, validation, construct definition, language testing

ACKNOWLEDGEMENTS

While writing this thesis, I have been a member of a number of professional and research communities, and I wish to thank my colleagues, friends, and supervisors for their help and support in the course of the work.

The Centre for Applied Language Studies in Jyväskylä has been my home base, and the sense of a collegial community during coffee breaks has given me a secure mooring both in good times and when the rest of life, including research, has made me feel insecure. I am grateful to my supervisors, Kari Sajavaara and Sauli Takala, for their unfailing support and continuing faith in me, and for their kindness to exert quiet and positive rather than voiced pressure on me to get this work finished. The whole CALS community helped me stay committed to reaching that goal.

The DIALANG project has formed a large part of my professional community in the past few years, and as a European test development project, it has provided me a virtual community for practical test development work, material sustenance in the form of pay for work done, and a circle of friends and colleagues throughout Europe. Special thanks are due to Neus Figueras for sunshine, good food and discussions both serious and entertaining, and to Charles Alderson for inspiring theoretical dialogues and warm personal talk, mostly about mountains and travels – the sources of many a smile. Furthermore, the discussions with Jayanti Banerjee that I have managed to have in the past few years have been personally and professionally rewarding.

One further testing community, that of the Finnish National Certificates, deserves my thanks for inspiring this work. In this context, I would like to acknowledge with gratitude Sauli Takala's key role in getting me involved in language testing and supporting my research interests. Discussions with Aila Määttä at the National Board of Education and with Anu Halvari and Mirja Tarnanen locally in Jyväskylä helped me define the area I was interested in.

I am greatly indebted to the external reviewers of this dissertation, professors Charles Alderson of the University of Lancaster and Pirjo Linnakylä of the University of Jyväskylä, who offered detailed and perceptive comments which made the work much better and clearer than it would otherwise have been. Its faults, of course, remain my own.

I am also grateful to the Centre for Applied Language Studies and the Nyyssönen Foundation for their financial support during the writing of this thesis.

I would never have stayed sane through this process without exercise. My thanks are due to Mirja Tarnanen, Peppi Taalas and Johanna Peltola for their company and for their patience in listening to long and convoluted monologues, which helped me clarify my thinking.

Finally, I am deeply grateful to my family and friends near and far who have given me material and emotional support that has carried me through this process.

Omistan tämän työn äidilleni ja isälleni.

WHAT DOES YOUR TEST MEASURE?
Construct definition in test development and validation

TABLE OF CONTENTS

CONCLUDING DISCUSSION

Appendices

List of tables

List of figures

1 INTRODUCTION

This thesis is about the development and validation of language tests. It focuses on the work of examination boards, whose responsibility it is to develop tests and prove their quality. As the reference to examination boards indicates, I concentrate on public, large-scale language tests. Such tests are visible and widely used, at least in their own contexts. The decisions made on the basis of these tests are often high-stakes, ie. they influence the life course of the examinees, regulating eg. the entrance to university. The tests need to be accurate and accountable, fit for their purpose, and accepted as useful by the test takers and score users. The requirements can be met by developing and validating the test in a professionally acceptable manner.

The test developers face a dilemma, however, especially concerning validation. The theoretical literature is not normally written from the point of view of test developers. Furthermore, theoretical developments over the past fifty years have introduced conceptual changes which, while making validation "more scientific" and thereby more credible and valuable, have also made it complex, abstract, and difficult to approach. "When do we do validation?" is a very real question to test development boards. So is "How do we do it?". The underlying questions here include "What is validation?", "What should we validate?", and "In terms of agenda for tomorrow and next week, what should we do?" In the course of the present thesis, I will address these questions.

One source of answers to the questions is to turn to theory. There are a number of theoretical works concerning test development and validation in educational and psychological measurement, including books and articles specifically concentrated on language testing. However, theoretical texts on test development tend to focus on test development and only mention validation in passing. Similarly, theoretical texts on validation tend to concentrate on validity and have little to say on test development. My aim is to combine the advice from both areas of theoretical writing and investigate ways of "doing validation" during test development.

Apart from consulting theoretical literature, new test development boards in search for guidelines for their work might also consult reports on the development and/or validation of an existing test, and this is the second approach I take in the present thesis. The individual development and validation reports (eg. Alderson and Clapham (eds.) 1992, Hasselgren 1998, Norton 1992, O'Loughlin 1997) are practical, but they naturally only

concern the test which they report on. Thus, they do not discuss the range of alternatives available to an examination board but only explain the solution which the particular board has chosen for developing and validating their test. Furthermore, some test development and validation reports are published as part studies rather than monographs: the series of research reports and technical reports concerning the Test of English as a Foreign Language (TOEFL) test (Educational Testing Service (ETS) 1999c) are a well known example of this. Such "distributed" reports are often written without an explicit framework of how each individual study or article is related to other reports and studies on the development and validation of the same test. The overall validation rationale does not necessarily arise as an issue. I consider the rationale a very important aspect of test developers' work, and this is a central object of interest in the present thesis. To analyse a range of existing practice, I will look at three cases where the tests serve the same purpose but, judging by the publicity information available on them, they are likely to represent different development rationales.

A fairly obvious source of advice for the developers of public, large-scale examinations is to consult professional standards for test development. Probably the best known set of standards is the *Standards for Educational and Psychological Testing*, published jointly by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). This document is currently in its sixth edition (AERA, APA and NCME 1999, henceforth AERA 1999), and it is widely consulted and referred to by language testers. The advice in this and other comparable standards (see eg. Alderson, Clapham and Wall 1995 for a summary of existing standards specifically aimed at language testers) is comprehensive and provides recommendations and advice for the development of effective and well-constructed tests which are supported by validity evidence. Measurement textbooks refer to recommendations in different standards and provide detailed advice for practitioners about how to achieve the standards. The most detailed and implementable advice in such textbooks concerns the psychometric quality of tests especially through statistical analyses of scores and score relationships. This is crucially important for the development of responsible testing practices, but to be able to explain the meaning of the scores, work on the theoretical definition of the construct(s) assessed in the test is also needed. There is much less advice available in theory for how to develop theory-based definitions of the construct assessed and how to use these definitions as

evidence of the quality of the test. In the present thesis, I will emphasize this aspect of test development and validation.

My interest in the theoretical construct definition arises from personal experience. I was involved in the development of a test in the early 1990s, and in the course of the development process, I stubbed my toe (painfully) on the problem of how to define the construct we were assessing and how to use the definition to improve the evolving examination system systematically. I read validity theory and consulted frameworks for test development. I found that I was expected to just do it. I was to define explicitly what was meant by the ability being tested, explain how the tasks implemented this ability and how I knew that the examinees were displaying it in their answers, and furthermore, I was to account for the criteria used to assess the ability. It was difficult. While I could see how to use the examination-in-making to develop the definitions, I was not sure how to use the evolving definitions to improve the examination or especially to prove its quality. It seemed clearly linked to validation: the definitions were words used for explaining what was intended to be assessed in the test. Yet I found no clear advice from theory or examples from practice on how to use this in validation. I did find advice for how to ensure the psychometric quality of the instrument, but our project did not have enough data yet for sophisticated statistical analyses. There was too much to do in the test development project to create a theoretical argument to support the use of the theoretical construct definition data in validation. Later when score data became available, statistical analyses were conducted and words were needed to explain the scores. We made some use of the characterisations and understandings of the construct assessed that had been developed in the course of the development work but I had a sense that more could have been done. In hindsight, it would have been possible to work on this if there had been a systematic record of the construct considerations that the development process had included. It would have helped if there had been a rationale or an example to follow which could have explained the benefits that this extra effort would bring. At the time I simply had an uneasy feeling that we could have done better, and I connected it with the earlier pain of not finding advice for how to do what I was clearly expected to do. Now, many years later, this dilemma finds its expression in the present thesis. In focusing on the theoretical construct definition in language test development and validation, I give this topic the attention I think it justly deserves.

1.1 Constructs, tests, and scores: theoretical and psychometric dimensions

Although much of published research in language testing is concerned with various aspects of test development and validation, the topic of my study has to my knowledge not been addressed before in the way I am doing it. This is partly because test development tends to be seen as a professional activity rather than a research topic, and partly because I have made a conceptual division between theoretical and psychometric dimensions in the important notion of *the construct*, which is not standard practice. I have done so to focus on activities in test development which I find important to address. Concentrated attention to the theoretical dimension can focus test developers' attention to what the test is testing through its tasks and assessment criteria, and thus help them build a coherent assessment system. This supports validation: if the challenges of current validity theory (eg. Messick 1989a) are taken seriously, questions about the nature of the skill assessed cannot be avoided. They are embedded in the notion of the construct together with the psychometric concerns of score dependability.

What is a construct? From a test development perspective, it is a concept (or a group of concepts) used to explain what the test tests. Anastasi (1986:4-5) defines constructs related to testing as "theoretical concepts of varying degrees of abstraction and generalizability which facilitate the understanding of empirical data." Chapelle (1998:33) defines them even more simply as meaningful interpretations of observed behaviour. In the context of testing, there is a continuum of increasing specification and concretisation from concepts through constructs to their operationalization in test tasks and examinee behaviours. Constructs are expressed in words, and the words used to say what a test score means identify the construct used in the interpretation, for instance "listening" or "overall proficiency." These, in turn, are related to theories of listening and proficiency, because constructs are not only associated with tests, they are also the building blocks of theory.

A key property of test-related constructs, according to Chapelle (1998:33-34), is that they should reflect performance consistency. This is required because only through performance consistency can researchers and testers move beyond observation of discrete instances to generalizations about what the observations indicate. To detect consistency, several observations of the same category of performance must be made. Together, the observation categories form the constructs assessed. It is possible, of course, that performance on some ability levels is consistently

inconsistent. However, if no consistency can be detected at any ability level, it would be difficult to argue that realisations of a construct are being observed. In such a case, the relationship between the construct theory and its operationalization in the tasks and observation categories has to be checked, and if the operationalization is acceptable, the theory probably requires revision, or at least it has not gained support from empirical trials.

If and when performance consistency is observed, there are still several alternatives for explaining what causes it, and as Chapelle (1998:34) points out, theorists differ on what they consider this to be. She follows Messick (1989a:22-30) to distinguish between three main groups of theorists as to what they consider to underlie performance consistency in tests. Trait theorists attribute consistency to characteristics of test takers, behaviorists attribute consistency to contextual factors, and interactionalists attribute it to traits, contextual features, and their interaction. Chapelle (1998:43) contends that there is strong support for the interactionalist perspective in current theory in language testing and second language acquisition. Within the interactionalist perspectives, there are differences between theorists about what kinds of constructs are considered important for explaining performance consistency. I will discuss this in more detail in Chapter 4.

A language test is an instrument which is used to express specific aspects of the examinees' language ability through scores. "Language ability" is usually the highest level construct in language tests, but no test can implement all of it. This is both because language ability is a very broad concept while tests are composed of a limited number of tasks administered over a limited period of time and because the definition of language ability in enough detail to enable test construction is not possible. Instead, through their tasks and assessment criteria, tests measure specific aspects of language ability. The construct definitions related to the tests should specify which aspects they are intended to measure.

The primary purpose of a language test is to *measure* the desired ability, although tests can have other functions as well, such as to promote learning (Bachman and Palmer 1996:19). To be able to measure, one of the bases of a test must be a continuum of ability. Different score levels on the continuum should denote different degrees or amounts of the ability assessed. Scores indicate the construct measured in a test, and since they are expressed as points or categories on an ability continuum, the accuracy and meaningfulness of the numerical scores are dependent on the measurement quality of the test. This is determined by the degree of consistency of scores and by the extent to which different scores can truly be said to denote different amounts of the ability assessed. This is why the

psychometric properties of tests and scores, primarily their reliability, are emphasized in all texts concerned with the development, validation, and use of tests.

While the dependability of the numerical values of scores is important, current standards of educational measurement state that it is equally important for test use that the developers are able to say what the test scores mean in terms of the abilities assessed. The plain numerical values of the scores are sufficient for score users only to the extent that the users want to rank examinees or sort them into categories according to whatever measurement definitions have been implemented in the test. If the users want to interpret the scores, verbal definitions are needed. Current standards promote such a wish in score users. Through this, they promote the sharing of power and responsibility for score interpretation between test providers and score users by making both score meaningfulness and score dependability equal technical criteria for test quality. In many contexts, such as diagnostic assessment in education, this has led to a situation where the more detailed the information that the test can provide, the more useful it might be considered. This must be qualified with the measurement proviso that the detailed information must be dependable, and with current psychometric means, dependability is easiest to show for general, test level measurements where variation in detail is not taken into account and may even be undesirable. While arguments for information potential promote multidimensional constructs, arguments for measurement quality promote generic constructs. Spolsky (1995) has expressed this consideration in an elegant contrast between the traditional or "humanist-scepticist descriptive approach" and the more modern "rationalist-empiricist measurement approach":

> In the course of my studies, I have come to believe that there is some value in each of these ideologies, and many problems with both of them. To oversimplify, with the traditional examination we think we know what we are assessing, but remain happily or unhappily uncertain about the accuracy or replicability of our assessment; with the modern examination, we are sure enough of our measurement, but are, or should be, uncertain as to what we have measured. Spolsky (1995:5)

To develop a good test, the developers need to pay attention to both the dependability and usefulness of the scores and the clarity, comprehensiveness, and usefulness of the verbal definitions of what the test is testing and what the scores mean. When they do, different types of activities are entailed and different questions are asked. The existing literature in testing and measurement provides comprehensive advice for the development of psychometric quality in tests. In the present thesis, I

investigate how work on the theoretical construct definition can support the content quality and interpretability of the test and its scores because I think that there is far less information and support available to test developers about it. However, since both are required, I will frequently come to the interface between the theoretical and the psychometric.

Although I have consciously limited psychometric discussion outside the present thesis, some definition about the interface is needed. The basic meaning of reliability is consistency of measurement; in other words, the degree to which the same result is obtained if a measurement is repeated. Consistency of measurement builds on the accuracy of the measuring instrument and the consistency of the procedures in repeated administrations. This is relatively easy to establish if the construct measured and the method of measurement are concrete and simple, while it becomes more difficult when the complexity of the construct or the measurement instrument increases. Because of the high value of simplicity, reliability can be contrasted with validity when this is defined as the comprehensiveness with which the measurement instrument reflects the construct. In the context of language testing, Davies (1990:50) states that "a completely reliable test would measure nothing; and a completely valid test would not measure." While this illustrates the contrast, it is hardly likely that either extreme would exist in reality. As many writers contend, both properties are continua, and the aim in developing a good test is to develop sufficient degrees of both. It is common to speak of tradeoffs between reliability and validity, although some of the current views of validity also see reliability as a concern encompassed in the broad concept of validity. I will discuss the broad definition in Chapter 3.

1.2 Research questions

The "real world" question that I investigate in the present thesis is: How can test development and validation be implemented in practice so that the principles of accountable measurement are followed? To address this topic systematically in a research sense, I analyse it from two main perspectives. These are the recommendations that are made in theoretical texts concerning this topic, on the one hand, and the reported actual practices of test development that are implemented by language testers, on the other. Each of the perspectives is analysed through specific research questions.

The theoretical texts which are particularly relevant to my research topic concern test development, validation, and construct definition. Since the texts currently available about language test development are primarily

addressed to test developers while texts about validation and construct definition are not, I will treat the literature on each of the topics separately so that I can tailor the specific questions to the nature of the literature. All the questions are joined by their aim of distilling advice to test developers about the nature of test development and validation work.

The existing literature about language test development is quite coherent in its approach; the authors tend to view test development procedurally and present principles of good practice in test development. My research questions focus on theory's advice to test developers, and they could all be preceded by the phrase "According to test development theory". The questions are:

- What steps does the process of language test development consist of?
- Which qualities should the developers pay attention to when developing the test?
- What should the test developers do about validation?

The existing literature on validation is usually not addressed directly to test developers. Thus, the questions I investigate are motivated by a wish to see the validity literature from a test development perspective. The questions I will investigate are:

- What is validation?
- What should test developers validate?
- How should test-related validation be implemented as a process?
- What is the role of construct definition in validation?

The existing literature on the definition of constructs for language tests provides a varied set of approaches to construct definition. I will discuss them under a sub-division into theoretical and empirical approaches. The questions I will investigate are:

- What is the nature of the constructs that the different approaches define?
- How are the constructs related to, and reflected in, the test instrument, the testing process, and the test scores?
- How can test developers use the different approaches to construct definition in test development and validation?

My analysis of different test development practices is based on reports and studies published about three tests that differ from each other in terms of the nature of their construct definitions. I will analyse the reports in the form of a case study using a set of detailed questions. I will present the

questions and the rationale for the case study design in Chapter 5. The questions that I will address at the level of cross-case comparison are:

- What are the similarities and differences between initial and operational test development and validation?
- How does the nature of the construct definition in a test influence the rationale followed and the quality criteria used in test development?
- How does the nature of the construct definition in a test influence the rationale followed, the questions asked and the results published on validation?
- How do the examples of realised practice in language test development and validation correspond to recommendations from theory?

In the conclusion to the thesis, I will bring together the answers to the research questions listed above and present directions for future research and practice in test development and validation.

1.3 Materials and methods

To give an overview of how the research questions are addressed in the present thesis, I will briefly summarise the materials and methods used in the study. The methods are discussed in more detail in the course of the thesis, particularly in Chapter 5.

The present thesis is a conceptual analysis of theory and practice in language test development and validation. It is based on analyses of two types of texts, theoretical texts about test development, validation and construct definition, and published articles and reports related to specific test development cases.

My purpose in analysing theoretical texts is to develop an understanding of test development and validation as processes from the point of view of test developers and specifically to clarify the role of the theoretical construct definition in them. This analysis constitutes Part One of the thesis. In Part Two, I will use the understanding that I have developed from the literature to construct a framework of analysis, which I will apply on three reported cases of test development and validation. The rationale for case selection will be discussed in Chapter 5. The materials concerning each case comprise all the reports and articles that have been published about the development and validation of the test during a development phase which I analyse. I will define the limits of each case at the beginning of the case

report, where I will also include a detailed outline of the materials used in the case.

The method I use in Part One of the thesis is conceptual analysis. One of the outcomes, a summary model which shows test development and validation as parallel processes, will be used to define the object of study in the multiple case study which constitutes Part Two of the thesis. I chose the case study method for the thesis because it supports an organized combination of theory and practice to analyze processes. I will analyze multiple cases because it allows me to investigate a range of actual practices in language testing and thus complement existing research which includes single-case reports and one survey of practices. The three cases that I will analyse represent different development rationales, especially in terms of the role of the construct definition. The purpose of the analysis is to find out how this influences the processes of test development and validation in terms of the questions asked, the data used, and the support developed for the quality of the test. I will discuss the design and implementation of the case study in Chapter 5 of the thesis.

1.4 Organisation of the thesis

The thesis has two main parts followed by a discussion. Part One analyses advice from theory and Part Two analyses current practice as it is reflected in reports about the development and validation of existing tests. The discussion draws together the advice from theory and the implications from practice in a renewed set of suggestions for practices in test development and validation, especially in terms of defining what is assessed.

Following this introduction, Part One consists of three chapters. In Chapter 2, a range of frameworks and quality recommendations for language test development is analysed and an outline of interlinked steps in test development is developed. An assessment is also made of the relationship between test development and validation as it is portrayed in test development theory. In Chapter 3, ways of conceptualising validity theory are discussed and issues that are particularly relevant for test developers are highlighted. The case is made that construct validation can begin early in the test development process if this stage is used for recording the evolving construct definitions that the development process implements. Chapter 4 discusses theoretical and empirical approaches to the definition of constructs for language tests. The results show a range of alternatives which test developers can use when they develop construct definitions for their tests and study the constructs actually assessed in them.

Part Two analyses reported practice in test development and validation. Its five chapters constitute a multiple case study. In Chapter 5, I will describe the design of the case study, discuss the considerations in case selection, and present the protocol that I follow in the analysis. Chapters 6 to 8 then report the results of the case analysis in accordance with the protocol. Chapter 9 discusses cross-case comparisons in terms of the research questions presented above. The extent to which the differences between the cases are due to the role of the construct definition is assessed and additional and alternative explanations are considered.

The final chapter, Chapter 10, takes a renewed look at the research questions from Part One in the light of the results of Part Two. The nature of test development and validation activities is discussed in terms of desirable goals and means of addressing them. Ways of defining constructs for language tests are discussed again and a set of dimensions for future analyses of test-related constructs is proposed. The limitations of the present study are discussed and suggestions are made for future research on test development, validation, and the definition of constructs assessed in language tests.

Part One
Recommendations from theory

2 THEORETICAL FRAMEWORKS FOR LANGUAGE TEST DEVELOPMENT

In this chapter, I will look into theory's advice on how language tests should be developed. I will investigate three questions:

- What steps does the process of language test development consist of?
- Which qualities should the developers pay attention to when developing the test?
- What should the test developers do about validation?

The purpose of this chapter is to introduce the world of language test development to the reader through the "best practice" or "advisable practice" presented in textbooks and theoretical articles. This world is what this thesis is about, so it is important to discuss its nature in some detail. At the same time, the discussion provides a summary of theory's advice to test developers. In order to give a comprehensive picture of this, I will analyse a relatively large number of texts. I will not analyse all textbooks and articles ever published, but I will include the most recent and most frequently quoted materials written about test development for language testing specialists and for language teachers. For comparison and contextualisation, I will also include two frameworks from educational measurement, of which language testing forms a part. This field has generally provided the professional standards quoted in language testing.

Regarding each book or article, I will briefly characterise the nature and purpose of the text, and then summarise what it has to say about the questions listed above. To conclude the analysis of each work, I will briefly list the distinctive features of the text in relation to the other texts I analyse. In the final section of this chapter, I will summarize the state of the art in test development, addressing the above questions again and drawing implications for later chapters in this thesis.

2.1 Best practice and realised practice: Alderson et al. (1995)

Alderson, Clapham and Wall's (1995) *Language Test Construction and Evaluation* describes and discusses the different stages of language test development. The book is intended for language teachers and other professionals who need to construct language tests, and for people who need to select tests developed by others (p. 1). The authors present what they consider to be best practice in language testing, but they also describe

realised practice through summarising and discussing the results of a survey of how British language testing boards develop their tests. The book offers illustrations of techniques, examples of products, outlines, and checklists for each of the stages to help intending testers or testing boards conduct their work according to the best practices of the field. The authors explain all the technical concepts that they use, and introduce basic principles of item analysis, so that readers are equipped to deal with more advanced textbooks, including psychometric ones, as and when necessary.

The approach to test development in Alderson et al. (1995) is practical and action-oriented. Regarding test specifications, for instance, the authors begin by explaining what they are and who needs them. They then give a long and detailed outline of possible contents for a specification, and an example of what an actual specification might look like. They explain that the writing of a specification normally begins from the description of the purpose for the test, the target audience, and the framework within which the test might be constructed, such as a linguistic theory or a needs analysis. Finally, they summarise British EFL examination boards' responses to a survey concerning their test specifications, and evaluate this against the 'best practice' described earlier in the chapter. The chapter finishes with a checklist of possible contents for a specification. The same pattern of presentation is followed in all the chapters which concern test development. The authors first explain the concept and then describe the activities which are involved in implementing it. Next, they summarize how a range of British examination boards report that they have done the activities concerned, and finally they summarize the advice in terms of best practice. In the concluding chapter to the book, the authors present the idea of professional standards as codes of practice, and review the different sets of standards and codes for test construction to which they have referred in earlier chapters.

2.1.1 View of test development

Alderson et al. (1995) divide their treatment of test development into nine chapters. Each of them describes a logical stage in the test development process, but the authors mention several times that the stages influence each other, so that the reality of test development is recursive rather than a linear progress from one stage to the next. The stages are: writing test specifications, item writing and moderation, pretesting and analysis, training examiners and administrators, monitoring examiner reliability, reporting scores and setting pass marks, validation, preparing post-test reports, and developing and improving tests based on feedback.

The authors begin with test specifications, which they consider a record of what is to be measured and exactly how this is to be done. The starting point for a test specification is the purpose of the test. In addition to this, according to Alderson et al. (1995:11-20), specifications can include a very wide range of definitions, including: a description of the test takers, a definition of the test level, a characterisation of the construct to be assessed, a description of suitable language course or textbook, the number of sections or papers in the test, the time for each section or paper, the weighting for each section or paper in calculating the final score, a definition of the target language situation, characterisations of the text types to be included, definitions of text length, a list of language skills and possibly language elements to be tested, the types of test tasks which can be included, sample instructions for each of the task types, criteria for marking, descriptions of typical performance at each level, description of what candidates at each level can do in the real world, sample papers or tasks, and samples of students' performance on tasks. This exhaustive list shows how detailed test specifications can be, but Alderson et al. (1995:10-11) also point out that specifications need not be this detailed for all user groups. The most detailed specifications are needed by test writers and test validators. Test users would probably find a version which concentrates on intended uses, score meanings, and empirical data on the scores most useful, whereas potential test takers would benefit from descriptions of the test content and sample papers and performances. Dedicated versions could also be written for teachers who prepare students for the test, for admissions officers, or for publishers of language textbooks. Furthermore, as mentioned above, test specifications are written and revised in iterative cycles with item writing and trialling so that modifications can be made once empirical evidence exists.

Regarding item writing and moderation, the authors emphasize the group work nature of this task. The result of item writing should be a set of tasks which test the right thing and which are free from technical errors. This can only be achieved through iterative rounds of task writing, commenting, and revision. The authors briefly discuss a range of task types and the most common technical problems with each of them. They refer to other resources which treat these more extensively, including Heaton (1988), Hughes (1989), and Weir (1988). They stress that editing committee members should always take the items they are judging as if they were test takers (p. 63), as only this will tell them about the procedural quality of the items they are judging.

In discussing pretesting and analysis, Alderson et al. (1995:73-103) make the point that empirical analysis is necessary regardless of how carefully the tasks have been developed, because only real trials will show how reliable the items are and how well they work in practice. The authors introduce and exemplify techniques for classical item analysis in an effort to demonstrate that empirical analysis is not difficult to do, and even small pools of data can yield meaningful results where the detection of defective items or keys is concerned. They emphasize, however, that the larger data pools that can be used in pre-testing the better, especially where selected response testing is concerned.

The authors devote a chapter to the training of examiners and administrators. Training helps ensure that different test administrators give the test in the same way and that different assessors, especially assessors of speaking and writing, work in a comparable manner. This is necessary for ensuring the reliability of the scores and the validity of the score interpretations (Alderson et al. 1995:105). The authors treat the monitoring of examiner reliability in a separate chapter, because the practical procedures for training and monitoring which they present are somewhat different, and both are needed for ensuring the quality of the assessment system. Rater training is focused on informing the individual raters about the assessment criteria used in the test, and qualifying the raters for their work by checking that they follow the agreed procedures. Administrator training mostly shares information on practices to be followed in giving the test. Monitoring is more a responsibility of the examination board. The authors discuss strategies which enable monitoring, such as sampling of rated performances for re-rating, use of reliability scripts, and double marking. They also discuss the concepts of intra- and inter-rater reliability, pointing out that "any agreement between examiners will be limited by the internal consistency of any and all examiners" (Alderson et al. 1995:135-136).

Regarding the reporting of scores, Alderson et al. (1995:150-154) discuss weighting, transformation, and score aggregation as ways of producing the scores which will be reported to test takers. Their recommendation regarding weighting, if it is to be used at all, is that the most reliable sections of the test should receive the most weight. They recommend transformation if test sections have different lengths but the developers want to give each section equal weighting in giving the final mark. Regarding reported scores, they discuss the alternatives of giving a single overall score or a profile score with separate results for important parts of the test. The authors contrast the testers' wish to provide more information with the administrators' need for a single score on which to

base decisions. This may mean that when scores are used, the score profiles will be neglected, whether they are reported or not. When they discuss the setting of pass marks, Alderson et al. (1995:158-159) make the point that examination boards should make decisions on these in the light of content considerations, and should report clearly how the pass marks have been set. Regarding validation, the authors particularly stress that validity cannot be taken on trust. Instead, examination boards must provide data and make arguments for the validity of their test for the use to which it is being put.

In discussing post-test reports, the authors stress the importance of standard reporting as an important form of regular feedback (Alderson et al. 1995:197-198). This can be useful for the testing board itself, for teachers who prepare students for the test, and for score users. Finally, in discussing the development and improvement of tests, the authors emphasize the need for continuous monitoring. Monitoring should focus on: the test; its implementation; results and their use; user comments; the test's relationship to theoretical and measurement technical development; and the practicality of the test (1995:218-225). Monitoring is necessary for providing proof that the test is working as intended, and for detecting any needs for change, major or minor. The authors give examples for how routine monitoring can help testing boards maintain a defensible assessment system, provided that there is then the ability and willingness to do something about any flaws that are detected. They discuss two kinds of examination change; one of constant small changes in response to monitoring information, and the other of major revisions, where a new test replaces the old version after a number of years of operation. Thus, for Alderson et al. (1995), test development is a recursive cycle which includes initial development, setting up of practices for administration, monitoring while the test is operational, and devising of new tests when feedback indicates that a new test is necessary.

2.1.2 Principles and quality criteria

Alderson et al. (1995) identify validity and reliability as the overarching principles for test development (1995:6). They define validity as "the extent to which a test measures what it is intended to measure", and consider it a property of test interpretations and uses, which are also influenced by test purpose (p. 6). The authors define reliability as "the extent to which scores are consistent", and consider this primarily a property of the test as a measuring instrument (p. 6). Validity and reliability act as quality criteria for all stages of test development. Regarding the relationship between the two, the authors begin by explaining that a test cannot be valid unless it is

reliable, because if a test does not measure something consistently, it cannot always be measuring it accurately (p. 187). However, a test can be reliable but not valid for the intended purpose: for instance, a multiple choice test on pronunciation can be highly reliable but it can nevertheless fail to identify students whose actual pronunciation is good or bad (p. 187). In practice, both properties are continua, and "it is commonplace to speak of a trade-off between the two – you maximise one at the expense of the other" (p. 187). They go on to explain that reliability and validity are intertwined rather than distinct as concepts, and that since there are different ways of calculating numerical indicators for each, a tester needs to know which values are being discussed in order to be able to interpret the values properly. The bottom line is that "since a test cannot be valid without being reliable, it is essential that tests are checked in as many ways as possible to ensure that they are reliable" (p. 188).

Alderson et al. (1995) also discuss activity-based principles for test development in each of their chapters. They recommend that test specifications should be developed and published (p. 9) and that the items should be developed and moderated carefully following the specifications and in collaboration with a moderation committee (pp. 69-70). Further, items should be pretested so that their empirical properties are known before operational use (p. 74). Examiners and administrators should be trained and their work monitored to ensure the consistency of examining procedures and hence the validity and reliability of the scores (p. 105, 115, 128). Regarding scoring and pass marks, the authors state that decisions on combining item-level marks to arrive at reported scores and on setting pass marks should be made empirically and rationally rather than arbitrarily (p. 159). Tests should be validated in as many ways as possible especially using different kinds of evidence (p. 171), and post-test reports should be written and published to promote accountability and in accordance with the legal and moral obligations of the examination board (p. 216). Finally, tests should be monitored on a routine basis at least for matters concerning test content, administration, training, and marking (p. 218).

2.1.3 View of validation

When introducing their discussion of validation, Alderson et al. (1995:170-194) explain that they will use the traditional language of dividing validity into types, although these are essentially "different 'methods' of assessing validity" (p. 171). They use three categories: internal, external, and construct validity, all of which can entail both logical and empirical analyses. Internal validity refers to features within a test, such as acceptability to candidates,

expert evaluations of content, and investigations of test-taker processing. External validity refers to a test's relationships to other tests or criteria, usually evaluated through patterns of correlations between test scores and numerical indicators for the measures that the test is being compared with. The authors note that the term criterion-related validity is often used for this concept, but since they use the term "criterion" in the sense of performance criterion, they selected a different term. Construct validity refers to a large array of investigations into what the test is testing. Depending on the research question, this can involve comparison with theory, internal correlations, comparisons with students' biodata and psychological characteristics, multitrait-multimethod analysis, or factor analysis. The authors discuss the kinds of research questions addressed with each of the methods, and refer interested readers to other sources which explain the more complex procedures in greater detail. When reporting the responses by the British examination boards on their validation practice, Alderson et al.'s (1995:192) conclusion is that the boards appear to conduct very few empirical validation studies. The range of possibilities in the questionnaire shows that the authors consider validation a process which extends from the beginning of test development into operational test use, but the practice reported in the boards' replies indicates that most of the British EFL boards engage in validation during initial test development only. Empirical studies once the test is operational are mostly limited to equating different test forms.

Alderson et al. (1995:11-21) also discuss validation in connection with test specifications. The authors make a distinction between three user groups for these: test writers, validators, and test users. They consider it important that validator specifications include as much detail as possible about the theory behind the test. Theories can be implicit or explicit, they explain, but language tests always implement some kind of beliefs and theoretical concepts about language proficiency, language learning, and language use. These are composed of psychological concepts, or *constructs*, and the relationships between them. Records of these beliefs in the specifications form an important basis for construct validation. Furthermore, according to the authors, the specifications should also specify the test developers' view of the relationship between the skills tested and the item types used in the test. Once the test begins to be administered and performance data accumulates, the hypothesized relationships can be investigated empirically.

2.1.4 Distinctive characteristics of the text

The treatment of test development in Alderson et al. (1995) is distinctively practical. The authors give concrete advice on how to develop tests and an operational introduction to important concepts and quality criteria. Compared with Bachman and Palmer (1996; see below), the emphasis on practicality means that the examples that Alderson et al. (1995) discuss are real and sometimes less than ideal. This approach illustrates how the theoretical concepts introduced in the book can be realised in actual practices and what might be done to improve less-than-ideal ones. This gives the readers a practical understanding of what it is that testing boards do. The only step that the authors do not treat in detail is the development of scoring criteria and assessment scales, but they refer to other sources that do. Another distinctive feature is an introduction to the best-known standards and codes of practice that are available in the world and that intending testing boards might want to follow.

2.2 Principles and procedures: Bachman and Palmer (1996)

Bachman and Palmer's (1996) *Language Testing in Practice* aims to "enable the reader to become competent in the design, development, and use of language tests" (p. 3). To the authors, such competence involves having a theoretically grounded and principled basis for the development and use of language tests, and skills to make their own judgements and decisions about usefulness in particular situations. The target audience is language teachers, members of examination boards, applied linguists, and graduate students; in other words, people who need to develop or choose language tests for teaching, certification, or research purposes.

The book is divided into three main parts. The first part contains the conceptual basics for language testing: a model of language ability, a framework of test task characteristics, and an approach to the evaluation of test usefulness, which the authors see as the most important consideration in test evaluation. The second part presents principles and templates for the stages of test development. The third part includes a set of illustrative examples, where the conceptual tools presented in the first two parts of the book are used in language testing projects. The first example in particular serves the illustration function, and the authors refer to it several times during the conceptual discussion in the first two parts of the book. Bachman and Palmer suggest (1996:14) that students studying language testing through this book can use the other examples to practice the

techniques suggested earlier in the book, or they can consider them as examples which can be extended and modified to suit different situations.

2.2.1 View of test development

Bachman and Palmer (1996) organize the process of test development conceptually into three stages: design, operationalization, and administration. They clarify: "We say 'conceptually' because the test development process is not strictly sequential in its implementation. In practice, although test development is generally linear, with development progressing from one stage to the next, the process is also an iterative one, in which the decisions that are made and the activities that are completed at one stage may lead us to reconsider and revise decisions, and repeat activities, that have been done at another stage." (Bachman and Palmer 1996:86).

The first stage, design, involves the description of the purpose of the test, the characterisation of the test takers, the definition of the test construct, and the creation of a plan for evaluating the usefulness of the test. The authors give detailed advice on what to specify and define, and a rationale for why this must be done (Bachman and Palmer 1996:97-115). The motivation for the contents of the definitions is the test purpose. The scores, which stand for language ability, are used in the social world to make decisions such as selection, placement, diagnosis, or progress grading. Each of these decisions requires different kinds of proof that the test is relevant and that the inference based on the scores is valid for the decisions to be made. Clear descriptions of test purpose, relevant types of language use, and intended test takers, provide an important basis for test developers to demonstrate this. The authors also explain that the initial descriptions provide an important baseline for subsequent studies on the impact of the test on society.

The construct definition, according to Bachman and Palmer (1996:116), serves three purposes: it is a basis for score use; it guides test development; and it enables test developers and validators to demonstrate the construct validity of the test. Regarding the definition of the construct to be measured, Bachman and Palmer (1996:75-76; 127-128) suggest that the definition be divided between a design statement, where the construct is defined abstractly, and operationalization, where the construct is defined in more concrete detail for each of the test tasks. The authors make a case that language 'skills' – that is, reading, writing, listening, and speaking – are not part of the abstract, theoretical construct, but instead belong to the more concrete, 'appearance' side of construct definitions, which is properly addressed at the operationalization stage.

The second stage, operationalization, encompasses the writing of specifications for the test tasks and the creation of the actual tasks based on these, the creation of a detailed blueprint for the whole test, the writing of instructions, and the development of the scoring method. This stage provides the transition from the abstract definitions at the design stage to the concrete reality of an examination. To make the transition methodical and accountable, Bachman and Palmer propose a template for task characteristics, a structure for the analysis of necessary parts of test and task instructions, and a discussion of rationales and principles for different scoring methods.

Bachman and Palmer (1996:171) propose that the test task is "the elemental unit of a language test", and so the central activity in operationalization is the writing of test tasks. To ensure that all relevant concerns are addressed in task writing, the authors propose that detailed specifications should be written for each of the tasks. Their template for what should be included in task specifications includes the following parts: the purpose of the task, which will be one of the overall purposes of the test; a definition of the construct to be measured in the task; the characteristics of the task setting in terms of the physical setting where the task is completed, the participants, and the time of the task in relation to other tasks in the test; time allotment; task instructions; textual and linguistic description of the characteristics of input, response, and relationship between input and response, in terms of channel, form, language, length, and organizational, pragmatic and topical characteristics; and scoring method (Bachman and Palmer 1996:172-173). Their example of this approach to task specification and writing is presented in Project 1 in Part 3 of the book. The format of the specification is a table. Each of the headings listed above is repeated in the first column, and in the second column, the characteristics of the task are described with a few phrases. The task itself is presented in an appendix to the Project. Moderation and revision are not discussed. This approach thus emphasizes task analysis more than the practical work of task development and revision.

Bachman and Palmer (1996) stress the importance of clear instructions and provide a framework for writing them. They recommend that a separate set of general instructions should be written for the test as a whole, but also that each of the parts or tasks of the test should be accompanied by specific instructions. The instructions should inform the test takers about: the test purpose; the abilities to be tested; the parts of the test and their relative importance; anything important about the testing procedures such as whether the test takers can move to the next part of the

test directly upon completing the previous part and whether they are allowed to leave early; and the scoring method to be used (Bachman and Palmer 1996:184-189). The motivation for such detailed instructions is that the test takers should know exactly what they are expected to do so that they can perform at their best. At the same time, the instructions should be simple enough so all test takers understand them, and short enough, so that the testing time is mostly spent on the tasks.

Bachman and Palmer (1996:193) begin their discussion of the development of scoring methods from the idea of the measurement process. The process consists of a theoretical definition of the construct, an operational definition of the construct, and the development of a way of quantifying responses to test tasks. The last step is implemented through decisions on the scoring method for the test. They identify two broad approaches: scoring based on the number of test tasks successfully completed, and rating the quality of performances on rating scales. In both approaches, quantification means "specifying the criteria for correctness, or the criteria by which the quality of the response is to be judged, and determining the procedures that will be used to arrive at the score" (Bachman and Palmer 1996:195). The authors discuss the rationales behind each of the scoring types, reminding readers that the decisions must be motivated by the test developers' view of what is being tested in the test, but they also remind them about the importance of training raters and monitoring their rating.

The third stage of test development, administration, involves the test developers in administering the test, monitoring the administration, collecting feedback, analyzing the scores, and archiving test tasks. The key aspect of this stage, according to Bachman and Palmer (1996:245), is the collection of empirical data during test administration in order to assess the qualities of test usefulness. This happens at two main points in time, during pre-testing, and during operational administration. The feedback provides information about the test takers' language ability and about how well the administration and the test tasks work. The first is used for reporting scores, and the latter for developing the system further.

2.2.2 Principles and quality criteria

The main quality criterion for test development activities, according to Bachman and Palmer (1996:17-18), is overall usefulness. This is a combination of reliability, construct validity, authenticity, interactiveness, impact, and practicality, and development is aimed at maximising the overall quality rather than the constituents individually. Bachman and Palmer

(1996:19) see reliability and validity as the essential *measurement* qualities of the test. These characteristics are of primary importance for tests because the most important purpose of language tests is to measure language ability. However, the authors suggest that it is important to consider tests in the larger societal or educational context in which they are used, and therefore, the other qualities are also important for the overall usefulness of the test. Authenticity and interactiveness are features of tasks, and indicate the extent to which test tasks resemble some of the non-test tasks to which test users want to generalise the scores. Impact relates to the test's relationship with the society that uses the scores, and entails evaluating the social effects of using the test. Practicality is a property of test implementation. Bachman and Palmer (1996:36) consider practicality important because this determines whether the test is going to be used at all: impractical tests will soon be abandoned. As tests are always used for specific purposes in specific contexts, the criteria for usefulness are context-specific. Bachman and Palmer (1996:134) argue that minimum levels for each of the qualities can and should be set, but that these levels cannot be set in the abstract. They must always be considered in relation to an individual testing situation.

2.2.3 View of validation

Bachman and Palmer's formula for test usefulness contains construct validity as one of the ingredients. They see validity as a property of the interpretations of test scores. In the test, the construct is verbally defined in the test blueprint, and operationally defined in the tasks which the test includes. Because the score interpretations are tied to a particular domain of generalization, the analysis of construct validity depends on the relationship between the properties of the test tasks and the properties of the domain into which generalizations are to be made and the areas of language ability engaged by test tasks on the one hand and non-test tasks of interest on the other. This is why, according to Bachman and Palmer (1996:22), both the authenticity of the test tasks and the degree to which they engage relevant language abilities in the test taker, something which they term "interactiveness", must be analysed to support a validity case.

Validation, according to Bachman and Palmer (1996:22), is an ongoing process. Validators must build "a logical case in support of a particular interpretation" and demonstrate through evidence that the interpretation is justified. They could show, for instance, that the test content is relevant to the intended interpretation and that each form covers a sufficient range of the intended content domain, that the scores from the test are related to other relevant indicators of the ability in which the test users

are interested, and that the scores are useful for the prediction of the test takers' performance on some test-external task which the score users are interested in predicting.

The procedures which Bachman and Palmer (1996:133-149) recommend for validation form part of the procedures which they suggest for the evaluation of test usefulness. Three stages are involved: setting a minimum acceptable level for construct validity while also paying attention to other usefulness concerns, evaluating construct validity logically, and gathering empirical evidence for validation. The minimum acceptable level cannot be a single statistical estimate but must rather be a definition of a mimimum set of evidence of different kinds to be provided in support of construct validity. They propose a set of questions for the logical evaluation of construct validity. The questions focus on the clarity and appropriateness of the construct definition, the appropriateness of the task characteristics with respect to the construct definition, and possible sources of bias in these characteristics. The questions about the task characteristics also cover the scoring procedures used. About evidence for construct validation, the authors state that this should be both qualitative and quantitative, and that the gathering of evidence should begin early on in the test development and continue into the administration stage. Their proposal for the kinds of data that this may involve includes "verbal descriptions of observations by test administrators, self-reports by test takers, interviews, and questionnaires . . . statistical analysis of numerical data, including test scores and scores on individual test tasks" (Bachman and Palmer 1996:149). The plan for usefulness, to be created during the first stage of test development, should include a plan for what kinds of data are to be gathered and when this will be done.

2.2.4 Distinctive characteristics of the text

Bachman and Palmer's (1996) approach to test development is detailed and thorough. They provide both theoretical foundations and concrete templates and procedures for language test development. They particularly concentrate on detailed planning so that the planning and operationalization stages of test development receive a great deal of emphasis in the book. For test delivery, they recommend procedures for ensuring comparability and comfort, but they do not discuss the practice of operational development or test maintenance. A distinctive feature of the book is Bachman and Palmer's overarching quality criterion of test usefulness. This drives the whole test development process, so that a plan for how to evaluate usefulness is developed at the very beginning of a test development process, and

evaluations of it are made throughout the development process. Since the concept is intimately tied to the purpose of the test and the foreseeable consequences of its use, test purpose is strongly emphasized in the text.

2.3 Principles and practice for teachers: Hughes (1989)

Hughes's (1989) *Testing for Language Teachers* is specifically targeted at teachers, and it concentrates on classroom assessment. In fourteen concise 6-15-page chapters, the book provides introductions to basic concepts in language testing, advice for implementation, and some concrete examples. Like Bachman and Palmer (1996), Hughes sets out to dispel myths about mysterious, powerful and threatening language tests by giving teachers an understanding of what language testing is about. In the first half of his book, Hughes discusses purposes and kinds of testing, validity, reliability, backwash, and test construction. In the second half, he discusses test techniques for overall language ability, speaking, reading, listening, grammar, and vocabulary. For each of these, Hughes outlines the kinds of skills or sub-skills which might be assessed, gives generic advice for the specification and writing of the tasks, lists possible assessment techniques, and comments on scoring. In a final chapter, Hughes briefly lists an ordered set of points to observe when tests are administered. He points out that although the list may seem tedious, it is useful and worth observing, because sloppy administration can endanger the reliability and validity of the results.

2.3.1 View of test development

Hughes's (1989:48) chapter on test construction presents "a set of general procedures", which he then illustrates with two examples. The first step is to describe the purpose of the test as clearly as possible. This is followed by writing specifications, which should define the content of the test, its format and timing, criterial levels of performance, and scoring procedures. Hughes instructs that the content of the test should be specified as fully as possible. However, he warns that too detailed a specification may go beyond what is currently known about the nature of language ability. He recommends that writers of specifications should stick to "those elements whose contribution [to a language skill] is fairly well established" in current theory (Hughes 1989:49). He suggests that when reading, writing, listening, or speaking are tested, it might be useful to define operations (the tasks that candidates must be able to carry out), types of text to be included in the tasks, addressees to whom the test takers are writing or speaking, and the kinds of topics that might be included in the tasks.

After the specifications have been written, the developers move to the writing of the actual test tasks. Hughes stresses that task writing should be teamwork, which involves drafting, commenting, rejections, and revisions. Once the tasks have been agreed on, similar drafting, discussion, and revision of the scoring key should follow. Finally, Hughes recommends that before the test is given for real, it should be pretested. He recognises that this may not always be possible, but this probably means that there will be some problems with the operational test. Whenever the teacher plans to re-use the test or some of its tasks, Hughes recommends that the problems which became apparent in administration or scoring be noted down, and that the test be analysed statistically. In an appendix, he introduces some very basic statistical techniques for the analysis of tests and items (Hughes 1989:155-164). These include descriptive statistics, split half reliability, and standard error of measurement for test level data, and item-test correlation, facility values, and distractor analysis for item level data. He concludes the appendix with a one-page conceptual introduction to item response theory.

2.3.2 Principles and quality criteria

The principles that Hughes discusses for language testing are validity, reliability, practicality, and backwash. He states that ”a test is said to be valid if it measures accurately what it is intended to measure” (Hughes 1989:22), and discusses reliability in terms of consistency of measurement and score dependability. Practicality is a matter of observing resource limitations, particularly time and money. However, practicality cannot rule test development on its own, but instead it must be evaluated together with the other principles (Hughes 1989:8, 47). Hughes stresses backwash or the effect of testing on teaching and learning (p. 1). Alderson et al. (1995) call this concept ‘washback’, and Bachman and Palmer (1996) discuss it under well-planned tests as a way to introduce positive backwash. Furthermore, he emphasizes the importance of knowing precisely what the *purpose* of the test is going to be.

2.3.3 View of validation

Hughes discusses validity in terms of content, criterion-related, construct, and face concerns (1989:22-28). Content validity builds on representativeness, and attention to this aspect ensures that the most important, rather than the easiest, targets of assessment are included in the test. Criterion-related validity means comparing the test scores against a criterion, usually through correlation. The main varieties are concurrent validity, which refers to relationships with other assessments obtained at the

same time, and predictive validity, which refers to the correlation between test scores and some indicators of future performance that the test is supposed to predict. Construct validation is a research activity which involves demonstrating that the test measures what it is supposed to measure. Hughes warns that construct validation is slow and the best ways for a teacher to deal with it are to keep up to date with theories about language proficiency and to test language abilities as directly as possible. By the latter he means that writing should be tested by making students write, spoken interaction by having students engage in spoken interaction. Hughes (1989:27) defines face validity as the appearance of validity, but emphasizes the meaningfulness of it being related to the acceptability of a test to test takers, teachers, and test users.

Hughes's discussion of how particular tests are validated is short and simple; what it entails is criterion-related validation. Thus, the validation of a placement test would consist of the proportion of students assigned to appropriate versus inappropriate classes; the validation of an achievement test means a comparison against teacher ratings (Hughes 1989:57). This is a narrow view of validation, but Hughes' motivation for this was probably that his book is intended for teachers, whom he apparently does not expect to think about developing proof about the conceptual relevance of their tests for the teaching and learning activities. As discussed above, his view of construct validation for teachers was that this involved a priori work, so that tests are on up-to-date theories of language use.

2.3.4 Distinctive characteristics of the text

Hughes's book is very clearly written for a target audience, classroom teachers. It introduces the most basic concepts in language testing and minimal procedures for implementing them in testing practice. Hughes emphasizes the benefits of group work in the writing of tests, and guides teachers towards thinking about the skills they are testing. He suggests that teachers can get support for this from theory as well as textbooks. His instructions for how test tasks are to be written cover the skills tested, possible task types, and the scoring procedures that the tasks involve.

2.4 Task development for teachers: Weir (1993)

Weir's (1993) *Understanding and Developing Language Tests* concentrates on the development and revision of test tasks. The book is organised according to tests of different skills: spoken interaction, reading, listening, and writing. A first chapter discusses general issues in test

construction and evaluation, and a concluding chapter summarises the thrust of the book and looks into the future of test development. Each of the skill-specific chapters first summarises research on the nature of the skill discussed, and then considers the nature of the test situation and the criteria relevant for assessing the skill. This is followed by a presentation of a wealth of task types, with discussion of what particular skills are being tested and *how* they are tested through this particular task type. Furthermore, Weir discusses the pros and cons of task types from the point of view of the teacher who needs to prepare all the task materials, implement and assess the tests, and produce results which are as valid and reliable as possible.

2.4.1 View of test development

Weir (1993) regards test development as one of the classroom teacher's professional activities which are aimed at supporting and enhancing learning. He argues that testing should be done *well* to achieve this aim while negative impact, such as adverse reactions to test tasks and harmful influence on teaching and learning habits, is avoided. He proposes that good testing can ensue if teachers think about what they want to test, know about the alternatives of how the different skills can be assessed, and work together by commenting on each other's draft tests and assessment criteria. He provides a framework for the planning and evaluation of test tasks. The framework covers three dimensions: the operations, ie. activities or skills, to be tested; the conditions of performance while the learners are taking the test; and the quality of student output, which refers to the assessment criteria to be used and the ideas about levels of language ability which underlie them.

The basic unit in testing on which Weir concentrates is the task. He wants to make teachers plan their test tasks carefully, so that they should serve a certain purpose which is motivated by the teacher's understanding of language skills. This begins by thinking and talking about *what* the teachers want the tasks to test, and then making reflective decisions about *how* these skills should be tested and how the performances are to be assessed. The idea is to create links from intention to implementation, so that the skills originally envisioned actually get tested. The means that Weir proposes is careful planning. In addition to discussing concrete examples of task types for each of the four skills, Weir provides generic guidelines for good test development under the headings of moderating tasks, moderating the mark scheme, and standardising marking.

Weir proposes that when moderating their own or other people's tasks, teachers should pay attention to the level of difficulty of both individual tasks and the test as a whole, making sure that each test as a whole covers a range of levels of difficulty. Furthermore, they should ensure that the tasks elicit an appropriate sample of the students' skills while avoiding of excessive overlap through including too many tasks on a narrow range of skills and omitting other skills altogether. In terms of technical accuracy, task reviewers should make sure that the tasks are easy to understand and that the questions are linguistically easier to comprehend than the actual task material. Moreover, they should assess the appropriacy of the total test time and the test layout. Finally, Weir points out that the task review process should help guard against bias arising from one-sided test techniques or cultural unfamiliarity of content. (Weir 1993:22-25.)

Weir suggests that when moderating the mark scheme, evaluators should check that the assessment guidelines define all the acceptable responses and their variations and that subjectivity is reduced as far as possible where assessment of spoken or written performances is concerned. Evaluators should also check that item weighting is justified on content grounds if weighting is used. He recommends that test developers should leave as little as possible of any calculation or summing activities to raters, because this is a potential extra source of error in sum scores. Reviewers should also check that the marking scheme is intelligible enough, so that a group of markers can be guaranteed to mark different sets of performances in the same way. His final recommendation for moderating the mark scheme concerns conceptual coherence in the assessment system: reviewers should check whether the skills required by the scoring operations, for instance spelling in open-ended reading comprehension tasks, are also what the scores are interpreted to mean. If the scores are only interpreted to convey information about reading, reviewers should perhaps recommend changes in the scoring procedures. (Weir 1993:25-26.)

As for the actual marking work, Weir states that standardisation is required to ensure uniformity of marking, so that any individual's score does not depend on who marked his or her performance. For Weir, standardisation means that the marking criteria are communicated to markers in such a way that they understand them, that trial assessments are conducted, that assessment procedures are reviewed, and that follow-up checks are conducted during each successive round of marking (1993:26-27).

2.4.2 Principles and quality criteria

The principles of good language testing for Weir (1993) are validity, reliability, and practicality. A test is valid if it tests "what the writer wants it to test" (Weir 1993:19). This presupposes that the test writer can be explicit about what the nature of the desired ability is. Weir argues for the development of theory-driven tests and supports this by always discussing existing theories at the beginning of his chapters on the assessment of language skills. His motivation for concentrating on test techniques is that if the tasks are flawed, it is possible that this threatens the validity of the test. He also briefly discusses authenticity under the heading of validity, making the point that although full replication of real life language use cannot be achieved in language tests, an attempt should be made to make language use in test tasks as life-like as possible within the constraints of reliability and practicality. The case he makes, albeit concisely, is very similar to Bachman and Palmer (1996).

Weir (1993:20) defines reliability as score dependability. This means that the test should give roughly the same results if it were given again, and more or less the same result whether the performances are assessed by rater A or rater B. Moreover, reliability is connected with the number of samples of student work that the test covers, since if the test only contains one task, it is difficult to judge whether the result can be generalized to just that task type, or whether it says something meaningful about the skill more broadly. Weir (1993:20-21) states that validity and reliability are interdependent in that known degrees of consistency of measurement are required for test scores to make any sense, but also that consistency without knowing what the test is testing is pointless. Furthermore, Weir connects reliability with the quality of the test items. Unclear instructions and poor items can make the test different for different candidates, thus affecting reliability. Similarly, sloppy administration can also introduce variation in the test which can influence the test scores, which makes consistency of administrative procedures partly a reliability concern.

For Weir, practicality is connected with cost effectiveness. In classroom contexts this concerns the teacher's and the students' time and effort in particular, but it also relates to practical resources such as paper or tape recorders, number of teaching hours that can be reserved for testing purposes, and availability of collegial support to comment on draft tasks (1993:21-22). Weir makes a strong plea that practicality should not outweigh validity of the authenticity-directness type. He states that although some task

types may be easier to administer and score, if the skills that they are measuring cannot be specified, the tests are not worth a great deal.

2.4.3 View of validation

Apart from a brief discussion of validity as a principle for test development, Weir (1993) does not discuss validation as an activity. However, he states that "validity is the starting point in test task design" (Weir 1993:20). Moreover, his whole approach to task design and revision is built on testing skills which the test developers can name and trying to guarantee that this is what actually gets tested during the assessment process. He does not discuss ways of providing proof that this is happening.

Weir 1990, to which Weir (1993) refers to as "the companion volume to this book" (1993:28), discusses construct, content, face, washback, and criterion-related validity. He makes a distinction between *a priori* and *a posteriori* construct validation. *A priori* construct validation involves a description of the theoretical construct that the test is intended to measure, and *a posteriori* validation entails statistical studies to investigate whether this is happening (Weir 1990:24). As for content validity, he argues that in classroom testing, given the restrictions on time and resources, *a priori* consideration of the content of test tasks is the most feasible validation procedure. He stresses the acceptability side of face validity and considers this important for the test to be effective, but joins others in warning that content and construct validities should not be sacrificed to acceptability. He defines "washback validity" or simply "washback" as the influence of the test on the teaching that precedes it. Finally, Weir sees criterion-related validity as "a quantitative and a posteriori concept" to determine the extent to which a test correlates with appropriate external criteria (Weir 1990:27). He argues against blind faith in criterion-related evidence, because the validity of the criterion can be questionable, and because it is possible that scores from a test correlate well with an external criterion, but the authors cannot say what the test is measuring. Weir (1990:29) suggests that an appropriate mix of validity evidence depends on the purpose of a particular test that is being validated. He also makes a case for a possible new combination of evaluation criteria that might be applied on communicative tests (Weir 1990:27). He suggests that in addition to content, construct, and washback work, systematic judgements could be gathered from students, teachers, and other users of the test on its perceived validity before the test ever gets administered. Only if the test passes this hurdle should "confirmatory a posteriori statistical analysis" be conducted, presumably against the posited factor structure of the test. The proposal is a reiteration

of Weir's emphasis on a combination of a priori and a posteriori work, and similar proposals, though perhaps with less emphasis on the theoretical/empirical division and the order in which studies should be conducted, are made eg. by Alderson et al. (1995), Bachman and Palmer (1996), and McNamara (1996).

2.4.4 Distinctive characteristics of the text

While Weir (1993) promotes principles very similar to those brought up in Bachman and Palmer (1996), and presents principles of task revision which largely cohere with Alderson et al. (1995), the distinctive characteristic of his book is his approach to test development through test tasks. Moreover, he is perhaps the most emphatic among the writers on test development about the need for test developers to specify in advance what skills their tasks are supposed to be testing, and try to make sure that this is what actually happens. However, he does not provide means for how to perform any checks.

2.5 Principles and processes: ALTE (1996)

The ALTE *Guide for Examiners* is a chapter-length document which describes "the practicalities of test construction that have to be recognised by any test designer in order to develop a 'good' test in the most general sense of the term" (ALTE 1996:3). The document is intended for test developers in general, and particularly for "those wishing to make use of the Council of Europe's 'Common European Framework of reference for language learning and teaching'" (ALTE 1996:1). The target audience is defined in this way because the work is a User's guide to the Council of Europe Framework (henceforth CEF) (Council of Europe 1996). The *Guide* briefly discusses the model of language ability described in the CEF (p. 2), and lists the chapters of the CEF which are particularly relevant for test developers (p. 3), but the bulk of the text describes practical procedures for test development which are needed in addition to the CEF if a test is to be constructed. The practical advice given in the *Guide* would apply equally well even if some other content basis was used for developing the test construct. The *Guide* identifies five phases in test development: planning, design, development, operation, and monitoring. The text focuses on the processes of test development rather than its products "in the belief that suitable products emerge from clear principles and well-designed processes rather than the other way round" (ALTE 1996:1).

2.5.1 View of test development

The discussion of the test development process in the ALTE *Guide for Examiners* (1996) begins with a brief overview of the five phases, and a discussion of the cyclical and iterative nature of the test development process. What this means is that the processes and products of each of the stages influence each other, and activities are recursive rather than linear. The most detailed instructions in the *Guide* concern the writing of test specifications, the procedures for organising the writing and revisions of test tasks, and issues in test writing from an individual test writer's point of view. In addition, the authors briefly discuss pretesting and construction of test forms from a pool of trialled tasks. The *Guide* finishes with a brief overview of various concerns in test evaluation.

According to the ALTE *Guide,* the test development process begins from a perception that a new test is necessary (1996:5). This leads to a planning phase where the needs of potential candidates and test users are analysed. This information is then used in the design phase for the purposes of the initial test specifications. The aim of test specifications, according to the ALTE *Guide,* is to "describe and discuss the appearance of the test and all aspects of its content, together with the considerations and constraints which affect this" (ALTE 1996:5). Sample materials should be written at the same time, so that user reactions can be gathered. In the development phase, the sample materials are trialled and performance data analysed. Based on this data, rating scales and mark schemes for the test are constructed. Feedback is gathered from those involved in the trialling and from potential test users for the revision and improvement of the system. At this stage, radical changes to any aspect of the test which appears to cause concern are still possible.

The ALTE *Guide* identifies a turning point in test development at the end of the "development" phase. At this point, the test specifications reach their final form, and test papers are constructed for operational use (ALTE 1996:6). The use of the "final" in connection with specifications conveys a strong emphasis on stabilisation and standardisation. The new test is then published and it begins to be used operationally. Once the test is introduced, it enters the operational phase. New items are written, vetted, edited, trialled, and revised according to the specifications that have now been set (ALTE 1996:13-20). In addition to operating the test, the test developers will also monitor the testing activities. This entails regular gathering of feedback from candidates and test users and routine analysis of candidate performances. The purpose of monitoring is to evaluate the

testing activities and assess any need for revision. If regular monitoring or more extensive studies of test use indicate that there is a need for a major revision, the cycle will begin again from a perceived need for a new test.

The model of operational test development that the *Guide* presents (ALTE 1996:13-20) follows similar lines to those presented in Alderson et al. (1995:40-72). Materials are commissioned, draft materials are vetted and edited, trial or pretest forms are constructed and given to trial participants, results are analysed, and materials reviewed to determine which items and tasks can be accepted as they are, which should be revised and re-trialled, and which rejected. The accepted materials are included in a materials bank, which is used in the construction of operational test forms. An additional aspect which the *Guide* discusses is official procedures which large-scale examinations might implement in commissioning test materials: details to give on the nature of the materials expected and the way they should be presented, details on deadlines and fees, forms which might be needed for the follow-up of large-scale examinations at the central examination board, such as those for indicating acceptance of commission and copyright issues (ALTE 1996:15-17). Such information is probably useful for new testing boards which are setting up their system of test development.

The instructions for individual item writers that the ALTE *Guide* provides (1996:23-32) are of a very practical nature. They discuss issues concerning text selection, authenticity, features affecting the difficulty of a text, choice of item types, task instructions, and scoring guides. Alderson et al. (1995) and Bachman and Palmer (1996) discuss the same issues under test specifications, and so this section of the ALTE *Guide* might usefully be regarded as guidelines for specification writing. The presentation chosen in the *Guide* may be motivated by the range of its target audience, which is not only examination boards but also classroom teachers, who do not necessarily write detailed specifications.

A perspective on test development which the other texts do not touch on and which the ALTE *Guide* discusses briefly is features of the institutional and political context into which the new test will be introduced. The *Guide* touches on this topic under the "constraints" part of their discussion of considerations and constraints that guide test development (1996:8-9). For ALTE, constraints include the expectations that may be placed on the new test by the society which is going to use it: acceptability to all who come into contact with the test or its scores, commensurability with existing curriculum objectives and classroom practice, and expectations in the relevant educational community on the nature of the test. Constraints also include more practical and material concerns such as the

intended level of difficulty of the test, and the availability of resources for developing and implementing it. The authors state that their list of constraints is not exhaustive, nor can it be made exhaustive in the abstract. Their point is that test developers must learn as much as possible about the practical context into which they are developing tests (ALTE 1996:9).

2.5.2 Principles and quality criteria

The ALTE *Guide* identifies four principal qualities of language tests: validity, reliability, impact, and practicality (1996:8). Validity is realised through measuring appropriately what the test claims to measure, reliability is defined as freedom from errors of measurement, impact as a call for inducing positive effects on individuals and classroom practice, and practicality as compatibility between the test's demand on resources and the resources available for developing and using it (ALTE 1996:8). To realise the principles, the *Guide* instructs that the first stage of test planning should involve an analysis where the considerations and constraints of the particular situation are assessed and their implications for test development noted. It discusses professional considerations, which involve specifying exactly what needs to be tested, and practical considerations, which involve arrangement details such as the size of the intended candidature and the number of examiners and rooms available. Their "constraints" focus also on the less easily tangible, political and emotional dimension of the fitting-in and acceptability of a new test.

2.5.3 View of validation

The ALTE *Guide* states that test validation "is an integral part" of their model of test development (1996:34). The authors do not define or describe validation beyond the initial description of it being about measuring appropriately what the test claims to measure (p. 8), but the statement about the integral status of validity is made under the heading of "evaluating tests", where the argument concentrates on consequences and test impact. Both during test development and while the test is used operationally, the *Guide* instructs, procedures must exist to "validate the test; evaluate the impact of the test; provide relevant information to test users; ensure that a high quality of service is maintained". One factor which the *Guide* suggests will help achieve this aim is finding and using the best available experts for all stages of test development, and training them for their work (p. 34). As for evaluating impact, the *Guide* suggests routine collection of data surrounding the test. This would involve candidate profiles, score user profiles, data on the teaching that underlies test participation, analyses of test-specific

preparation materials, public perception of effect, impressions of students, test-takers, teachers, and parents, and impressions of members of society outside education (p. 35).

2.5.4 Distinctive characteristics of the text

The ALTE *Guide for Examiners* (1996) provides a practically oriented set of instructions for test developers. It describes the institutional-organisational side of test development practice, and gives professional guidelines to help actual test construction. In a chapter-length treatment there is not much space to go into any of the topics in great detail, but the text gives a practically oriented overview. The distinction between the initial development of a test system and the development of tests to set specifications while the test is operational is made very strongly in this text. This is in slight contrast to other texts such as Alderson et al. (1995), where specifications are never considered "final" but always potentially changeable. The reason for the difference may be the brevity of the ALTE *Guide*, since the other texts also regard specifications relatively stable in relation to a published test, so that substantial changes in specifications are only possible in connection with examination revisions.

2.6 Developing performance assessments: McNamara (1995)

McNamara's (1996) *Measuring Second Language Performance* has two main aims: critical examination of the theoretical bases of performance assessment and introduction of multi-faceted Rasch measurement, which the author presents as a useful research tool for performance assessment data. He discusses the development of performance assessments in one chapter, and the validation of such assessments in a main section in the first content chapter of the book. McNamara's focus is on performance assessment, specifically the assessment of language abilities in occupational settings.

McNamara identifies two key characteristics for performance assessment. Firstly, the assessment situation involves a performance process, and secondly, this necessitates a qualitative judgement process, through which the test performances are converted into numbers. McNamara (1996:9, 86) presents a model of the testing process with several components: the candidate; the assessment instrument, which can include tasks and interlocutors; the test performance; the rater; the scale(s); and the rating. Any stage in the testing process involves an interaction between two or more of the components. From a test development and validation point

of view, he argues, this means that specific attention has to be given to the nature of the test tasks and the complexity of assessment. The argument is relevant for present-day language testing, because most tests now include some sections that ask the participants to produce an extended piece of writing, speech, or spoken interaction, which will have to be assessed through the kinds of judgement procedures that McNamara describes.

2.6.1 View of test development

McNamara presents test development as a staged process. He distinguishes ten stages, presenting them first in an ordered list, and then exemplifying the process through a summary of how the Occupational English Test (OET) was developed. McNamara describes the OET as ”an Australian government test of ESL for health professionals” (1996:98), in other words, a test of workplace English.

McNamara suggests that test development begins from the stating of the test rationale. He formulates this as a question “*who* wants to know *what* about *whom*, for what *purpose*?” (1996:92). When McNamara reports on this step about the OET, he uses the title “test background”, which is logical, because the responses to the question above provide a contextualisation for the test. The next step, according to McNamara (1996:94), involves an assessment of resources and implementation constraints, so that the development project is practical. He points out that the development and use of performance assessments is costly. After the resource assessment, a broad definition is developed about the test content. In the case of second language performance tests, this can involve consultation with expert informants, literature search, job analysis and workplace observation, and collection and examination of samples of language from the workplace. McNamara argues (1996:100) that in spite of Messick’s (1989a:17) view that content validity is not really validity at all because it is a property of the test rather than of the scores or score interpretations, content definition is very important for second language performance assessments. His list of empirical means, above, shows that content definition can be based on data from the real life language use situation even if further work with the data is required to turn it into functional test tasks. The point is actually fairly compatible with Messick’s view, which will be discussed in more detail in Chapter 3. What both authors consider important is the influence of the test tasks on the skills reflected in the test performances, which is important for the construct assessed.

The domain definition is used to inform the writing of test specifications and the writing of actual test tasks. McNamara sees the specifications as a set of detailed procedures which the writers of successive versions of the test are to follow (1996:96). They must also include definitions of scoring standards and procedures, development of rating scales, guidelines for the selection, training and (re-)accreditation of raters, and decisions on the reporting of the results. Next, a pilot version of the test is trialled and feedback gathered from trial subjects. In order to rate the performances, raters need to be selected and trained, and the data then analysed. The feedback gathered from participants, and the information from data analyses, will be used for revising the test materials and specifications in the next stage of test development. Finally, the minimum acceptable performance standards are set with the help of expert informants, and the test is ready to go operational. McNamara describes the procedure they used in setting the minimum standards for the OET. Since the main purpose of the test was to select those applicants whose language ability was sufficient to sustain them through medical familiarisation in an Australian hospital, a group of ten doctors was asked to rate 20 candidates on the speaking part of the test to find the minimum level. McNamara (1996:111) reports a high degree of agreement both within the group of Australian doctors and between the medical staff and ESL practitioners. Once the test is operational, McNamara mentions two main test development activities: implementation and monitoring. Monitoring should focus on the quality of test implementation, eg. the performance of raters and test reliability. Simultaneously, the empirical validation of the test proceeds.

2.6.2 Principles and quality criteria

McNamara (1996) does not directly discuss principles and quality criteria for test development. However, in the course of his book, he discusses the different arguments in favour of performance tests, the importance of theories of performance, and ways of modelling performance assessment. In so doing, he raises the values of representativeness and realism of test tasks and performances, generalizability, and construct validity. None of these concepts is simple for McNamara, and he argues for the need of both theory and empirical evidence to support the asserted value of performance assessment in these respects. His principal thrust is that the *assessment* in performance assessment requires more attention than it has received thus far. The means he proposes for providing empirical evidence about the

quality of the judgements is item response theory (IRT), particularly multi-faceted Rasch measurement (McNamara 1996:117-119).

2.6.3 View of validation

McNamara sees validation as an activity which begins during test development, and continues into the operational use of the test. The main question, for him, is "*how* and *how well* we can generalize from the test performance to the criterion behaviour. *How* involves decisions of test design; *how well* is an empirical matter which can be investigated on the basis of *data* from trial or operational administrations of the test, supplemented if appropriate with additional data, for example, on the *predictive* power of the test" (McNamara 1996:15-16; emphasis in the original). He discusses content validation, predictive validity, and washback or impact, and quotes Messick's (1989a, 1994) view that construct validity is the central concern in all validation. McNamara (1996:19) points out that an important way in which a construct definition appears in a test is its assessment criteria, whether or not the developers make explicit reference to a skill construct elsewhere in the test. However, he also argues for the need to tie language performance tests to a broad theory of language abilities, other abilities involved in language use, and their manifestation in actual communication. He contrasts this with total reliance on operational definitions of language ability, which he sees to underlie much of occupationally related language performance tests (McNamara 1996:8-9, 48-49). He finds support in Messick that a theoretical grounding help test developers articulate a theoretical rationale for inferences made on the basis of test scores, which is important in addition to empirical evidence. I will return to McNamara's discussion of language performance constructs in Chapter 4.

McNamara (1996:16) refers to Weir's (1988) contrast between *a priori* and *a posteriori* validation, but interprets this as a division between validation work prior to the publication of the test and validation work after it, rather than with a distinction between theoretical and empirical work on the development of the test. Implicit in McNamara's interpretation is a view that the publication of a test is somehow conceptually important for the test development process. However, McNamara does not discuss this in detail. When he describes validation work during initial test development, he stresses the importance of domain definition and the development of rating criteria. These feed into, and form part of, the test specification, and he states that "the development of test specifications is not a simple and mechanical procedure, but involves the test developer in facing a number of

complex issues of test validity" (McNamara 1996:96). Unfortunately, he does not explain this statement any further. At the same time, he also stresses the importance of empirical validation work: "The main validation research will be carried out in the field trials that precede the operational introduction of a test, or on operational versions of the test" (McNamara 1996:20).

2.6.4 Distinctive characteristics of the text

McNamara's text is very clearly focused on performance testing. The main difference between his text and other authors is his emphasis on the complexity of assessment. He discusses the link between assessment scales and construct definitions at some length, pointing out an area which has received little attention in language testing thus far. He distinguishes between 'strong' and 'weak' second language performance assessment, where the 'strong' version uses real life performance criteria for judging task performance, whereas 'weak' performance tests limit assessment to the features of *language* that the performance represents. His point is that whereas language tests tend to be on the weak side, we cannot guarantee that task completion features do *not* influence the assessments made. He does not offer immediate solutions, but he does give clear and emphatic expression to the problem. Similarly, he opens a discussion of what influences the performance in performance assessments, especially on tests of spoken interaction, since the speaking test event is an interaction between the candidate, the interlocutor, and the test task. This characterisation of the event is relevant for test development because it has fundamental implications for what the test scores mean. Test takers and score users tend to think that they reflect the ability of the test taker, and when important decisions are concerned, the onus is on the test developer to show to what degree this is the case.

2.7 Specification-centred initial test development: Lynch & Davidson (1994)

Lynch and Davidson's (1994) article about criterion-referenced language test development (CRLTD) promotes the writing of detailed test specifications as a way of strengthening the relationship between teaching and testing. The authors' approach is tailored for language teachers especially, but they state that it can be applied in other institutional contexts as well. Lynch and Davidson (like Davidson and Lynch 1993) propose a workshop approach to the writing of specifications and test items, where

the participants in the workshops are teachers who teach students for the test. The activities at the workshops cover the beginning stages of test development from the definition of the purpose of the test to the delivery of revised items for piloting. The authors mention piloting and finalisation of the test for operational use, but the approach that they promote does not encompass these stages of test development.

2.7.1 View of test development

Lynch and Davidson's (1994) criterion-referenced language test development builds on co-operation between teachers, who progress iteratively through a five-step process of writing and refining specifications and test tasks. The model is based on Popham's (1978, 1981) view of what a test specification should contain. This is: identifying information for the specification, a general statement of what is to be tested, a detailed description of the task that the student will encounter, a detailed description of what the student will have to do to respond to the task, a sample item, and any additional material which task writers will need to construct a relevant item (Lynch and Davidson 1994:731).

Lynch and Davidson have developed a framework for workshops which helps teachers write specifications and tasks according to Popham's model. Groups of participating teachers first define the mandate, or the motivation for the test, then write an initial specification for each item, or if source texts are used, then also for each source text, then write a sample item, and give the resulting initial specifications to another group. Each group then follows the specifications they have received and write items to fit them. As the final step, the whole group reconvenes to discuss the specifications and the items written so far. Feedback from participants leads to a need for revision in the specifications, the items written so far, and possibly also in the original definition of the test mandate. The process is started again, and iterations can continue as long as there is time available, or until the test developer is happy to start trialling the items produced. (Lynch and Davidson 1994:732.)

Lynch and Davidson (1994:732) argue that their detailed and iterative process of initial test development can help teachers develop a better understanding of their curriculum objectives and help them link the objectives with assessment activities. This can lead to *reverse washback*, or influence from teaching to testing (Lynch and Davidson 1994:737), bringing instruction and assessment into closer alignment through changed assessment. They seem to argue that the process they advocate is different from normal, psychometrically driven testing board practice, but the

processes of test development described in Alderson et al. (1995), Bachman and Palmer (1996), and ALTE (1996) at least are very similar to the one proposed by Lynch and Davidson. The authors also argue that detailed specifications and item refinement through the steps that they propose reduce the effort needed for psychometrically driven test refinement through trialling and item analysis, though they do not articulate the rationale in detail.

Lynch and Davidson's (1994) figure of test development identifies a starting point at defining the mandate for the test, followed by five stages: selecting the skills to be tested, writing specifications, writing items or tasks, assembling the test for piloting, and finalising the operational measure. Their CRLTD process encompasses the first three of the test development stages. The authors acknowledge the need to pilot items and monitor operational tests (Lynch and Davidson 1994:741), but these activities are outside the CRLTD process that they advocate.

2.7.2 Principles and quality criteria

Lynch and Davidson (1994) do not explicitly discuss quality criteria to be used to judge tests, but their whole approach is built on the value of good correspondence between teaching and assessment, something like content validity. They state that clearly defined constructs are necessary for a valid test (Lynch and Davidson 1994:730) and argue that detailed specifications provide evidence of validity because of the clear link that they make between the specification and the instructional goals. Other than this, they do not discuss validation; their focus is on the activities of the CRLTD process.

2.7.3 Distinctive characteristics of the text

Lynch and Davidson's (1994) article focuses on the initial stages of test development only. It argues for the value of detailed test specifications and a close link between teaching and assessment. It mentions validation in passing, but in doing so, it makes the point fairly strongly that detailed test specifications when developed through the CRLTD process to align with teaching provide evidence for a test's validity.

2.8 Options for educational test developers: Millman and Greene (1989)

Millman and Greene's chapter on test development in the third edition of *Educational Measurement* (Linn (ed.) 1989) concentrates on the

specification and development of tests of achievement and ability. The authors explicitly state that the chapter is aimed at professional test constructors, not classroom teachers (Millman and Greene 1989:335). Their goal is to discuss the range of different options available to test developers, rather than to give procedural guidelines for a standard test development process. It is this perspective of different purposes and options that motivates the inclusion of this text in the present overview. In procedural terms, the stages of test development that Millman and Greene cover are the same as those in the texts reviewed already.

2.8.1 View of test development

Millman and Greene's (1989) discussion is organised according to logical steps in test development. They begin with test purposes, then they discuss the possible contents of test specifications, followed by concerns in item development, item evaluation and trialling, selection of items for potential inclusion in tests, and assembly of test forms. The authors emphasize that test planning is fundamentally iterative, so that the stages influence each other.

Millman and Greene (1989:335) point out that the "most important step in educational test development is to delineate the purpose of the test." Their categorisation of purposes is different from many others, because it is not organised by the kinds of educational decisions that are to be made on the basis of the test, such as placement, diagnosis, and selection or initial evaluation, formative evaluation, and summative evaluation. Tests of achievement and ability are difficult to categorise according to such criteria, because while they should ostensibly belong to different categories, they share many functional purposes. Therefore, Millman and Greene (1989:336-337) categorise tests by the type of inference that will be made on the basis of the results. They distinguish between three domains and three types of inference. The domains are curricular, cognitive, and future criterion setting, and the types of inference are description of individual examinees' attainments, mastery decisions, and description of performance for a group or system. The curricular domain is further subdivided into domain inferences before instruction, during instruction, and after instruction. Each of the cells in the ensuing matrix identifies a set of test purposes with similar characteristics, such as diagnosis (description of individuals' attainments during instruction), program evaluation (description of performance for a group or system after instruction), certification (mastery decisions about a cognitive domain), or selection (mastery decision in relation to a future criterion setting). The authors' point is that the types and domains of

inferences have strong implications for features like test content and length, the kinds of items to be included, and criteria for evaluating items (p. 338). For instance, if inferences are drawn about individuals' abilities, each test form has to be representative of the domain and comparable to other individuals' test forms. If the inferences are closely related to an instructional program, the possible question types may be limited to those used in the instructional setting and the abilities and skills assessed to those of the curriculum objectives. If the inferences are related to a cognitive domain, individual instructional programs should not so strongly influence the definition of the skills assessed or the range of possible item types. Instead, such tests should be closely related to theoretical conceptualizations of mental abilities. If inferences do not concern individuals but instructional programs, comparability between different test forms answered by examinees is not an issue, but the study design as a whole should cover both content that the program focuses on and content in which it may be weaker. Millman and Greene's categorisation of test purposes is most useful for educational tests of achievement and ability. In the context of the present thesis, its benefit is the focus on the inferences drawn from the scores. Yet even in this categorisation, the tests that I will examine in Part Two of the thesis belong to two categories. Language tests used as admission criteria for university studies must be based on a theoretical conceptualization of the necessary language ability rather than on any curriculum specifications, but because of the selection function, they also refer to a future criterion setting. I will return to the issue of purpose in Part Two of the thesis.

Test specifications, according to Millman and Greene, should define test content, item types and psychometric characteristics, scoring criteria and procedures, and number of items to be developed (1989:338). Their discussion of alternatives for test content is thorough. It starts from the definition of the sources of test content, such as curricula or theories of ability, and the authors suggest that the content specification can be clarified especially through a characterisation of high performance in the domain being tested, for example through stating what experts can do compared with novices, or a characterisation of differences in strategies or knowledge structures between experts. The content definition should also make it clear whether the test construct is uni- or multidimensional, in correspondence with the curriculum or other source which the test is intended to operationalize. Furthermore, the authors point out that the content specification is influenced by the type of intended score interpretation, whether it will be domain- or norm-referenced. Domain-referenced scores

require clear specification of each sub-component in the domain, whereas norm-referenced scores rather require a clear definition of the main component(s) of the construct assessed (Millman and Greene 1989:341-342). If both kinds of inferences will be made, both content definition concerns will need to be addressed in the specifications, and balances be struck between broad and detailed content domain specifications, discrimination and content validity as criteria for item selection, and the appropriate rules for distribution of test content within each test form.

Millman and Greene (1989:343-345) also discuss the specification alternatives for scoring at some length. They contrast the relative ease of right-wrong scoring decisions with the potential for more detailed feedback if partial credit scoring is used. They recommend partial credit scoring especially for situations where feedback is desired on examinees' strengths and weaknesses. In terms of performance assessment, they discuss componential (or analytic) scoring, which they consider the most appropriate for multidimensional content specifications, and holistic scoring, which is the most suitable for unidimensional content. Both kinds of scoring require clear specification of proficiency at different levels, and careful development and quality assurance of the assessment process through the training of judges and regular monitoring of their work. They then discuss weighting, which they consider in the light of validly reflecting the content definition in the test. Weighting of content coverage, Millman and Greene explain, can be done by developing different numbers of items for different content areas according to their importance, and by applying weights that regulate how much importance individual items or groups of items have for the final score. This may require careful analysis of item statistics within content-motivated subsets of items to check that all relevant subsets contribute appropriately to the total score. Provision for such procedures should be made in the test specifications.

In the area of item writing, Millman and Greene (1989:349-351) discuss the continuum from the freedom of creative artists operationalizing a theoretical construct to almost mechanical rule-governed item generation according to detailed specifications. They finish by defending fairly detailed and prescriptive instructions for item writers, because it is easier to specify the principles by which such items have been created, and thus analyse their content.

Millman and Greene (1989:354) divide item evaluation activities into two broad categories: those where the content and format of items is judged against a set of criteria, and those where examinee data from item tryouts is used to evaluate item performance. They advocate the use of both

methodologies and the combination of their information when items are selected for operational tests and when test forms are constructed. They list item-content criteria as item accuracy and communicability; suitability of the item as judged against the content specification in terms of difficulty, importance and perceived bias; conformity to specifications; relevance to real-world tasks; and in educational contexts, opportunity to learn (Millman and Greene 1989:354-362). The criteria based on item response data that they discuss are item difficulty, discrimination, indexes based on subset-motivated patterns of item responses, and distractor analysis. They do not discuss Item Response Theory methods because these are discussed elsewhere in *Educational Measurement* (Linn (ed.) 1989). When introducing the performance data based criteria, the authors discuss important considerations in item tryout design, namely how an appropriate sample of examinees is acquired, how the number of items to be trialled is determined, and how test developers can decide a strategy for item tryout that corresponds with the practical setting in which the test is being developed. The alternatives that they discuss are using experimental items as the operational test, embedding items within operational tests, and arranging a separate tryout. Each strategy has its advantages and drawbacks, which the test developers need to tackle when they know how trialling will be done in their case. In addition to main trials, Millman and Greene (1989:356) recommend a small-scale preliminary tryout before the main trials to weed out gross flaws in instructions and task wordings.

2.8.2 Principles and quality criteria

In the introduction to their chapter, Millman and Greene state (1989:335) that although they "appreciate the importance of such factors as the cost, the consequences of an incorrect decision or inference, and the political and organizational milieu in which test planning and development take place", they will confine themselves to technical matters in test development. They do not list the principles that they promote in a straightforward list, but their discussion emphasizes good planning, coherence and quality assurance, especially through validity and reliability.

2.8.3 View of validation

Millman and Greene (1989) do not explicitly discuss validation in their chapter. However, throughout their text they treat validity as one of the criteria guiding test development. This is particularly apparent in their treatment of test specifications and through these, all concerns in test development which are related to the construct to be assessed. They state

that "the major function of [test specifications] is, quite simply, to enhance the ultimate validity of test-score inferences. Derived directly from the designated purpose of the test, the specification of test attributes provides a guide to subsequent item development, tryout, evaluation, selection, and assembly. This direct grounding of developmental activities in test purpose helps to insure the congruence between intended and actual test-score inferences and, thus, the validity of the latter." (Millman and Greene 1989:338.) They particularly use validity as a criterion in discussing definitions of test content, which for them encompasses construct definition, in judging item quality against content specifications, and in analyzing the appropriate weighting of each content area for scoring the test.

2.8.4 Distinctive characteristics of the text

Millman and Greene's chapter discusses the traditional phases of test development in the context of educational testing of achievement and ability. Its specific feature is the discussion of the alternatives that test developers have at each stage: sometimes a range that all test developers have to choose from, sometimes the practicalities of how different decision making purposes influence the activities undertaken at the same stage. None of the alternatives is perfect, but when a range of them are presented, it is easy for test developers to compare benefits and drawbacks. Another distinctive feature of Millman and Greene's chapter is its emphasis on the content/construct definition as a core for the whole test development process.

2.9 Test development in *Standards for educational and psychological testing* (AERA 1999)

The *Standards for educational and psychological testing* (American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) 1999:2) state that the purpose of the *Standards* is "to provide criteria for the evaluation of tests, testing practices, and the effects of test use." The *Standards* is perhaps the best known and most referred-to set of criteria for evaluating educational and psychological tests. Although the document is American, it is well respected in other parts of the world as well. The current *Standards* is the sixth revised edition of guidelines for test construction and use from the three sponsoring organisations, the first having been produced separately by APA and AERA in the 1950s. The target audience of the *Standards* is all professional test developers. The

document codifies a set of practices which the educational and psychological measurement community views as a desirable standard. The format of the standards is prescriptive, but there are no formal enforcement mechanisms; professional honour should compel test developers to follow them.

The new *Standards for educational and psychological testing* (AERA 1999) includes six chapters on test construction, evaluation, and documentation, one each for: validity; reliability; test development and revision; scales, norms, and score comparability; test administration, scoring, and reporting; and test documents. Each of the chapters first discusses general concerns related to its topic, and then presents and, where necessary, explains the standards related to it. The the introductory texts have been expanded from previous versions; their purpose is to educate future test developers and users and help all readers understand the standards related to each topic. The first two chapters concern standards for the key measurement criteria in the evaluation of tests, validity and reliability, and the last four contain standards for the stages and products of the test development process. Furthermore, Part Two of the *Standards* includes four chapters on fairness issues. There is some overlap in the scope of the chapters, and the chapter on test development and revision states that "issues bearing on validity, reliability, and fairness are interwoven within the stages of test development" (AERA 1999:37).

2.9.1 View of test development

The *Standards* identifies four main steps in test development: "(a) delineation of the purpose(s) of the test and the scope of the construct or the extent of the domain to be measured; (b) development and evaluation of the test specifications; (c) development, field testing, evaluation, and selection of items and scoring guides and procedures; and (d) assembly and evaluation of the test for operational use" (AERA 1999:37). It also states that the development activities are not always sequential but that "there is often a subtle interplay" between the stages, so that the writing of items and scoring rubrics clarifies the definition of the construct. Furthermore, the *Standards* emphasizes the idea that the rationale for a test is strengthened when both logical/theoretical evidence in the form of the framework and empirical evidence from item development and test construction are available to support the interpretations of test scores (AERA 1999:41).

The aim of the first step of test development, according to the *Standards* (AERA 1999:37) is to extend the original statement of purpose into a detailed framework for the test to be developed. The framework

”delineates the aspects (e.g., content, skills, processes, and diagnostic features) of the construct or domain to be measured”, and guides all subsequent test evaluation. The specifications, then, detail ”the format of items, tasks, or questions; the response format or conditions for responding; and the type of scoring procedures. The test specifications may also include such factors as time restrictions, characteristics of the intended population of test takers, and procedures for administration” (AERA 1999:38). Specifications are written to guide all subsequent test development activities, and they should be written for all kinds of assessments, including portfolios and other performance assessments.

The *Standards* points out that specifications must define the nature of the items to be written to some detail, including the number of response alternatives to be included in selected response items and explicit scoring criteria for constructed-response items (p. 38). The document identifies two main types of scoring for extended performances: analytic scoring where performances are given a number of scores for different features in the performance as well as an overall score, and holistic scoring where the same features might be observed, but only one overall score is given. The readers are told that analytic scoring suits diagnostic assessment and the description of the strengths and weaknesses of learners, while holistic scoring is appropriate for purposes where an overall score is needed and for skills which consist of complex and highly interrelated subskills. (AERA 1999:38-39.)

The *Standards* states (AERA 1999:39) that when actual items and scoring rubrics begin to be written, a participatory approach may be used where practitioners or teachers are actively involved in the development work. The participants should be experts, however, in that they should be very familiar with the domain, able to apply the scoring rubrics, and know the characteristics of the target population of test takers. Experts may also be involved in item review procedures, which can be used in quality control in addition to pilot testing. Such review usually concerns content quality, clarity or lack of ambiguity, and possibly sensitivity to issues such as gender or cultural differences.

In the final step of initial test development, the items are assembled into test forms, or item pools are created for an adaptive test. Here, the responsibility of the test developer is to ensure that “the items selected for the test meet the requirements of the test specifications.” According to the *Standards* (1999:39), item selection may be guided by criteria such as content quality and scope, appropriateness for intended population, difficulty, and discrimination. Similarly, the test developer must make sure

”that the scoring procedures are consistent with the purpose(s) of the test and facilitate meaningful score interpretation.”

For the purposes of score interpretation, the *Standards* makes the point (AERA et al. 1999:39-40) that the nature of the intended score interpretation influences the range of criteria used in the selection of items for the test. If the score interpretation is to be norm-referenced, item difficulty, discrimination, and inter-item correlations may be particularly important, because good discrimination among test takers at all points of the scale is important. If absolute, or criterion-referenced, score interpretations are intended, adequate representation of the relevant domain is very important ”even if many of the items are relatively easy or nondiscriminating” (AERA 1999:40). If cut scores are needed in score interpretation, discrimination is particularly important around the cut scores.

The actual standards for good practice in test development (see AERA 1999:43-48) require careful documentation of all test development procedures. Test developers should document the test framework, specifications, assessment criteria, intended uses of the test, and the procedures used to develop and review these. Any trialling and standard setting activities should also be documented in detail. Assessors should be trained and training and qualification procedures documented, administration instructions clearly presented and justified, and the public should be informed about the nature of the test and its intended uses in sufficient detail to ensure appropriate use of the test.

The *Standards* chapter on scales, norms, and score comparability (AERA 1999:49-60) presents rationales and standards which are directly related to score interpretation and score use. The text is relevant for the test development process especially in the sense that the planning and gathering of evidence to create reporting scales for the test, establish norms, and make decisions on mechanisms by which score comparability between forms will be ensured, has to start during test development. The strategies planned will be implemented throughout the operational phase of the test because new items and test forms will always require scaling, calibration, and equation. The chapter also explains what the assessment system must be like to ensure that scores from different forms can be compared and equated. This is not possible if different versions measure different constructs, there are distinct differences in reliability or in overall test difficulty between forms, the time limits or other administration conditions are different between the different forms, or the test forms are designed to different specifications. Furthermore, the chapter advises the readers that the establishment of cut scores, ie. points on the reporting scale which

distinguish between different categories of ability, is always partly a matter of judgement. The procedures used to establish the cut scores during test development, and the qualifications of the people who take part in the procedure should be carefully documented, so that the standard setting procedures can be reviewed and repeated if necessary.

For test administration and scoring, the *Standards* makes the point that the procedures for these activities given in the test documentation must be followed to ensure the usefulness and interpretability of the test scores (AERA 1999:61). If this is not done, the comparability of the scores and the fairness of the assessment system for individual test takers are endangered. This is also why it is important that the test documentation includes such instructions. The chapter on supporting documentation for tests (AERA 1999:67-70) lists the following features which a test's documentation should specify: "the nature of the test; its intended use; the processes involved in the test's development; technical information related to scoring, interpretation, and evidence of validity and reliability; scaling and norming if appropriate to the instrument; and guidelines for test administration and interpretation". The documentation should be clear, complete, accurate, and current, and it should be "available to qualified individuals as appropriate." Test users will need the documentation to evaluate the quality of the test and its appropriacy for their needs.

2.9.2 Principles and quality criteria

The *Standards* (1999) does not explicitly discuss principles and quality criteria for test development. The chapter on test development refers to validity, reliability, and fairness issues, and the standards for test development encourage careful documentation of all stages of test development, and the presentation of both theoretical rationales and empirical evidence to support cases for intended score interpretation and use. Validity is portrayed as the overarching concern in test development, focusing on score interpretations which are entailed by proposed uses of tests; reliability is related to consistency of measurement; and fairness in terms of test quality is envisioned to mean that the test should not contain deficiencies which cause the score interpretations to be different for identifiable groups of test takers, nor should the test documentation allow for the test to be administered or its scores used in such a way as to disadvantage identifiable groups of test takers.

2.9.3 View of validation

The *Standards* considers the validation process to be about ”accumulating evidence to provide a sound scientific basis for the proposed score interpretations” (AERA et al. 1999:9). Validation is focused on score interpretations, not on test instruments, and when scores are used for more than one purpose, each of the intended interpretations requires its own validity case.

Validation and documentation underlying test development are related in that validation should start from the test framework, which is one of the first documents that a test development body should draft. The test framework contains a definition of the construct, ie., “the knowledge, skills, abilities, processes, or characteristics to be assessed. The framework indicates how this representation of the construct is to be distinguished from other constructs and how it should relate to other variables. The conceptual framework is partially shaped by the ways in which test scores will be used.” (AERA 1999:9.)

The aim of validation, according to the *Standards* (1999:9), is to provide ”a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use.” Since the test’s framework is the starting point for validation, and the framework is influenced by the purpose of the test, test purpose has implications for test development and evaluation. Validation continues into the operational use of tests, and all evidence accumulated when a test is being offered is potentially relevant for old and new validity cases.

The *Standards* points out that validation involves careful attention to possible distortions of score meaning (AERA 1999:10). Such distortions may happen because the construct is inadequately represented by the test, or perhaps because some of the test methods turn out to have a significant effect on scores. “That is, the process of validation may lead to revisions in the test, the conceptual framework of the test, or both. The revised test would then need validation.”

The evidence that test development activities offer for validation is based on the documents related to test development, especially the test framework and specifications, and the items contained in test forms. The primary method of providing such data, as described in the *Standards* (1999:11-12) is expert judgement. Experts can be asked to analyse the relationship between a test’s content and the construct it is intended to measure, as defined in the test framework. If the test has been developed on the basis of a content domain specification, the items or score patterns can

be judged against this document to assess how well each test form represents the specification. Similarly, experts can be asked to judge the quality and representativeness of items against specifications. Furthermore, as the *Standards* points out, expert panels can identify potential unfairness in the review of test construct or content domain definitions.

2.9.4 Distinctive characteristics of the text

The *Standards for educational and psychological testing* (AERA et al. 1999) provides thorough documentation of professional standards for test development. The language is exhortatory, but effort has clearly been made to make the wordings clear and the standards comprehensible. The introduction to each chapter supports the comprehensibility and makes the *Standards* educative reading. The target audience is professional test developers, and the expectation is that the measurement or assessment procedure to which the standards are applied is relatively formal and standardised, as can be gleaned from references to alternate forms, standardised administration procedures, and different kinds of supporting documentation.

2.10 State of the art in test development

In this chapter, I have discussed theory's recommendations for test development practice. I have analysed nine textbooks and articles to review what they say about the test development process, the quality criteria which should be observed in developing a test, and the instructions which they give about the validation process. In the sections below, I will answer the three questions I presented at the beginning of this chapter and discuss the results and their implications in the context of the present thesis.

2.10.1 Consensus on the stages of test development

As has become evident in the course of the present chapter, theorists writing for various audiences about test development present a largely shared view of the logical stages which the process involves. Test developers thus have a secure basis in theoretical literature which they can use when they start to plan the test development activities.

The development process begins from the purpose and scope of the test, since this guides all subsequent stages of test development. This should be written down so that it can be referred to whenever needed. Next, the test specifications are written, explaining in detail what it is that will be tested and how it will be done. This is followed by the writing of test items

and assessment scales or scoring criteria. Further, procedures for the assessment of the performances and the administration of the test have to be developed, and if assessors are needed, they have to be trained and qualified. The test items are then piloted and the performance data analysed to evaluate the items and choose the most appropriate of them to construct tests. Administration and assessment procedures are also evaluated and revised if necessary. The result of this test development process is a test form which is ready for operational use.

In addition to the consensus on these stages, the theorists who write about test development are unanimous that test development is iterative and recursive as an activity. This means that the activities do not proceed in a linear fashion but are cyclical, and that the products of all of the stages influence each other.

2.10.2 Features particular for the development of formal examinations

The frameworks analysed in this chapter concern a range of language testing contexts. Some concentrate on features that are common to all contexts (e.g. Bachman and Palmer 1996), some focus specifically on teaching (e.g. Hughes 1989, Weir 1993), and some specifically on large-scale testing programmes (e.g. Alderson, Clapham and Wall 1995, ALTE 1996, Millman and Greene 1989). Since I will analyse the development and validation reports that concern large-scale testing programmes in Part Two of this study, it makes sense to briefly list the features of test development that are specific to such contexts.

The term "large-scale testing programme" is used to refer to tests that are fairly stable formal entities, developed and maintained by testing boards over considerable periods of time. They tend to serve relatively stable social purposes, such as the examination of whether prospective non-native speaker students have sufficient language ability for university study. The tests to be discussed in Part Two of the present thesis were developed for this purpose. The aim in using the scores as admission criteria is to ensure that admission decisions are made on an equal basis across different groups of applicants and across time. To ensure such comparability, different forms of the test have to be comparable and the test always has to be administered under the same conditions. This is achieved through careful initial development of an examination, standardization of the procedures used when new test forms are developed, standardization of administration, and careful monitoring.

The institutional difference between formal examinations and informal ones is that formal examinations involve an examination board. The more far-reaching the decisions that are going to be based on the test scores, the more possible it is that the examination board is called upon to defend the quality of the test, either in public discussion or legally in court. The examination board thus has a legal-institutional motivation to follow generally accepted professional procedures and record evidence which can be used to defend the examination. Documents about the standardised features of the examination will also probably be needed by the examination board simply to ensure that everybody involved in the examining process will know how to act. While the examination board is responsible for the standardisation and for policy decisions such as when a test revision is needed, the different types of activities involved in developing and maintaining a test will probably be implemented by smaller working groups. Thus, a range of test centres will be responsible for administering the test under standardized conditions, a group of test developers will be responsible for developing new test items, and when the decision is made for a new or revised test to be developed, the work will probably be seconded to a working party or a test development project. Both the policies of the examination board and the activities of the people involved in developing and maintaining the test are important for achieving an understanding of how a test is developed and validated.

As concerns large-scale examinations, the publication of the test constitutes a distinctive turning point in the test development activities. Prior to this point, all the test development activities can be summarised under the heading of product development. This involves the writing, trialling, and revision of the test, the scoring procedure, and all the documentation required for the administration of the test, as well as the training and qualification of all the personnel needed in the administration. Changes in one of the components of the system can lead to changes in the other components. The end of the initial development phase is marked by a standardisation process. The outline of the test and the procedures involved in its administration are set, and the specifications reach their final form (see eg. ALTE 1996). The test is then published and the official administrations begin. Test development is an ongoing concern even after the publication of the test, but the aim is changed from the improvement of the test system and its optimization to the maintenance of the system as it has been agreed it should stand, and the creation of items and tests comparable to the existing ones. The examination board will monitor any need for change or improvement, but if such need is established, a new test development

project will probably be created while the old form of the examination will still continue to be administered and used (see eg. Alderson et al. 1995). With the availability of empirical evidence from official administrations, validation should continue as long as the test scores are used.

2.10.3 Principles to guide test development

The frameworks of test development discussed above are similar in that validity and reliability are considered to be the most important measurement principles in test development. On a general level, the validity concern in test development is seen to be whether the test is actually testing what it is supposed to be testing. Validity is regarded as a property of score interpretations, and thus related to test purpose. Reliability is related to the consistency of the test scores, for instance between administrations or between raters. Reliability and validity are related, and both are required for the right thing to be tested with sufficient consistency for the purposes of the test.

Furthermore, the writers on test development tend to mention practicality as an important principle, although some authors (eg. Bachman and Palmer 1996, Hughes 1989, Weir 1993) account for it in more formal detail than some others (eg. Alderson et al. 1995, McNamara 1996, Millman and Greene 1989). Practicality involves checking what resources are available for the development and implementation of the test in terms of money, personnel, and time, and the production of the best possible test for those resources. Otherwise the test will not be practical and will not be used.

Bachman and Palmer's (1996) view of the principles that guide test development is slightly different from the other writers' in that they make the point very strongly that the main quality criterion for test development is overall usefulness. The other authors also mention several desirable qualities and discuss the importance of striking a balance between them; the addition that Bachman and Palmer make is that they place the main emphasis clearly on the utility of the test for its intended purposes. This is emphasized by the name of the key quality, "usefulness". In the context of formal examinations, it is likely that the decisions about the weights and minimum acceptable levels that the individual criteria receive are made at the overall policy level, probably by the examination board. Such decisions are undoubtedly guided by the test purpose so that in high-stakes contexts the reliability and validity of tests are likely to be highly emphasized, but they also reflect the values and beliefs of the decision makers.

The principles for good practice promoted by test development theorists are motivated by a desire to provide accountable measurement. The aims could be summarised in the following list:

- to measure the right thing
- to measure consistently
- to measure economically
- to provide comparable scores across administrations
- to provide positive impact and avoid negative consequences
- to provide accountable professional service

Much of the discussion around the principles and practice of test development is about the definition and operationalization of the construct to be measured. I will summarise the current discussion on the construct definitions used in language testing ventures in Chapter 4; at this stage I want to point out, as a summary of the texts discussed in the present chapter, that when language testers talk about principles and best practice for test development, they emphasize the importance of recording what the test should be testing. They urge test developers to check at every stage that this is actually what is assessed, but the only means that they offer for doing so are based on conscious planning and self-monitoring. The theorists clearly and unanimously recommend that the construct should be specified in the test specifications. There should also be detailed rules for its operationalization in the test tasks and assessment criteria. Missing from the discussion, however, is the consideration of whether and how the construct definition should be used to defend the quality of the examination in validation. Specifications are often confidential to testing boards, and the monitoring of whether and how their intent is realised in actual test forms is also not a common topic of publication. Such publications would be one obvious implied outcome from the theorists' advice, especially since it is possible that examination boards already do most of the work. The missing link is publication and, possibly arising from this, public discussion of how valuable such evidence is for validation.

2.10.4 Relationship between test development and validation

The test development frameworks take test development and validation as intertwined activities. The *Standards* (AERA 1999), for instance, states that test development *interweaves* issues bearing on validity, fairness, reliability, norming, and test administration, and that validity concerns are addressed in the chapter on test development as and when required. However, validity and the other issues are also accorded a chapter of their own in the

Standards, and the relationship between test development and validation is not spelled out in detail. In the standards for test development, validation is mentioned specifically in relation to score interpretation, outlining and assessment of the appropriacy of test content, and empirically based selection of items for use in the test. The most frequent instruction concerning test development is for the developers to *document* both the process and the products. The use of this documentation to defend the quality of the test is not discussed. Such practice would require that test developers begin to publish development documentation, and that test evaluators include its publication as a quality criterion when they evaluate tests. Possible evaluation criteria could include comprehensiveness and comprehensibility. On the basis of such documentation, connections might be developed towards applied linguistic research in two areas, construct definition and the checking of quality in operationalization.

Alderson et al. (1995) do not explicitly address the relationship between test development and validation, but they bring up validation particularly strongly in connection with test specifications, test content, and the ways in which the performances are assessed. Their chapter on validity discusses internal, external, and construct validation and emphasizes both theoretical and empirical work to examine validity. Bachman and Palmer's (1996) formula for test usefulness contains construct validity as one of the ingredients. They discuss reliability and validity (1996:19) as the main *measurement* qualities to be observed and relate construct validity concerns particularly to the construct definition and the characteristics of the test tasks (p. 21). They too make a distinction between theoretical and empirical investigations, and call for both. McNamara (1996:97-112) discusses content selection, development of assessment scales, and rater training aspects of test development as particularly relevant to validation. This view emphasizes the "initial development" part of test development, which McNamara considers important especially in the context of performance assessment. Weir (1993:19-20) considers validity to be related mostly to definitions and decisions on what to test.

It seems fair to conclude that according to test development theorists, test development and validation intermesh at the scientific basis for score meaning, which is the theory of the construct(s) or abilities behind the test. Such theories are reflected in the processes and products of the functioning assessment system, and their nature can be investigated by recording the rationales and products of test development. This documentation is necessary for test development, and it also provides the foundation for all subsequent validation activities. Exactly how the connection with validation

works is not discussed in the texts, although it is implied that these construct-related questions can guide the test's validation plan by raising questions to be investigated. The test development theorists also stress, however, that validation must involve empirical work on operational data from the test, and this work can only really begin after the test begins to be used operationally. Empirical validation is therefore an important concern after the test is published.

In the next chapter, I will review the state of the art in validity theory and the advice that this theory offers to test developers about validation. As part of the review, I will investigate how validity theory sees the relationship between validation and test development.

3 VALIDATION IN LANGUAGE TEST DEVELOPMENT

In this chapter I will analyse validity theory in educational measurement from the point of view of language test developers. I will address the questions I listed in the introduction to the present thesis:

- What is validation?
- What should test developers validate?
- How should test-related validation be implemented as a process?
- What is the role of construct definition in validation?

Validation is a broad and abstract topic, and it is not typical of theorists to talk about it as an operational process, certainly not one which is intertwined with test development. Therefore, it is not possible for me to treat the validity literature in the same way that I discussed frameworks for test development in the previous chapter; I cannot list ten different validation textbooks and analyse their advice for test developers. Instead, I will describe the different ways in which validity and validation are conceptualised in theoretical texts. The aim is to describe how test-related validation could or should be done according to validity theorists.

I will begin with a nutshell definition of the modern concept of validity and a quick overview of the historical development of validity theory. The historical view is one framework that test developers can use when they try to conceptualise a validity text that they are reading. This will be followed by a more detailed discussion of the current concept of validity and an overview of the current areas of theoretical debate, which provide further frameworks for conceptual analysis. Next, I will consider the advice from theory for conducting validation studies. I will conclude the chapter with a discussion of those issues in validity theory that are the most central for the present thesis, namely the status of the test in test validity and the role of construct theory in validation. While these are not "hot issues" in current validity theory, they are central practical problems for test developers.

3.1 Validity in a nutshell

Validity is a fundamental concept in the philosophy of educational measurement. It is concerned with correctness, truth, and worth, which are characteristics that mankind has always found interesting. In everyday speech, validity is used as a criterion for the adequacy or soundness of reasoning or statements. In educational measurement, validity deals with the meaning of the measure.

The meaning of the measure defines a broad area of interest. As Anastasi (1986:3) points out, "almost any information gathered in the process of developing or using a test is relevant to its validity. It is relevant in the sense that it contributes to our understanding of what the test measures." Messick's (1989a:24) outline of validation methods is equally broad: "Test validation embraces all of the experimental, statistical, and philosophical means by which hypotheses and scientific theories are evaluated."

In a landmark chapter, Messick (1989a:13) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment." The new Standards for Educational and Psychological Testing (AERA 1999:9) largely paraphrases this definition in less technical language. It defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed by specific uses of tests."

The two definitions above reflect the key features of modern validity theory in educational measurement. Validation is considered to be evaluation (eg. Cronbach 1988, 1989; Messick 1989a), that is, an informed human judgement made on the basis of multiple lines of evidence. The validity of something is evaluated, it cannot be exhaustively defined in lawlike statements or logical or statistical formulae. Thus, the current definition implies that validity cannot exist independently of the person who makes the validity evaluation. The current view is that validity is a unitary concept. This contrasts with the 1970s view that validity was divisible into rather independent, interchangeable types: content, construct, and criterion validities. The definition also indicates that validity is a matter of degree rather than an all or none matter. The fact that both theory and empirical evidence are mentioned implies that both are needed to create a balanced validity argument. Finally, the current view is that the object of validation is not the test or even the score. It is the interpretations of the scores in specific situations of score use. This means that validation is situation-specific and that it concerns both test developers and score users.

Cronbach (1988:3) summarises the work that validators have to do as "activities that clarify for a relevant community what a measurement means, and the limitations of each interpretation." The formulation is elegant and simple. However, the action that it implies is rather challenging for the developers of language tests. In this context, "what a measurement means" refers to the test's definition of particular aspects of language ability. The "relevant community" includes test takers and score users, whose different

vested interests in the interpretation of the score belong to the practical world of qualification and decision-making. It also includes the scientific community, which asks for precise definitions of the abstract concept of (foreign) language ability. Everyone needs explanations and evidence, and if these are to be believable, they must fit the framework of the examination system and address the community's questions. The challenge lies in how to do all this. Different answers are provided by the different stages in the development of validity theory.

3.2 Early developments in the history of validity theory

The history of current validity theory begins in the late 1800s with the birth of objective testing. With more than a hundred years of development, validity theory has evolved considerably, but at first the development was slow. The standard definition of validity in the first half of the twentieth century was that it was the extent to which a test "measures what it purports to measure" (Garrett 1937:324). This was chiefly judged in terms of how well a test predicted the criterion that it was used to predict, and it was operationalized as a correlation coefficient between the test score and the criterion value. According to Angoff (1988:19), validation work was "characteristically pragmatic and empirical, even atheoretical, and validity data were generally developed to justify a claim that a test was useful for some particular purpose."

One reason why validity theory developed slowly initially was that validity was not considered a problematic theoretical concept. It was a technical quality related to test use. Since educational tests were mostly used to predict performance on some criterion, validation simply required that the test correlated with the criterion. Bingham (1937:214), for instance, defined validity as the correlation of scores on a test "with some other objective measure of that which the test is used to measure". Guilford (1946:429) expressed this in even more radical terms: "In a very general sense, a test is valid for anything with which it correlates." The meaning of the scores was not the main focus of interest in validation; the usefulness of a test to predict a criterion was.

There were some tests, however, for which it was difficult to find an external criterion against which to compare them. These included achievement and proficiency tests, which measured the level of skill that an individual had acquired. In his review of the early history of validity, Angoff (1988:22) explains that measurement experts such as Rulon (1946) considered these tests valid by definition. They were their own criterion, and

all the validity evidence that was needed for them was a review by subject matter experts to confirm that the content of the test was representative of the domain of skill being measured. In the first edition of his *Essentials of Psychological Testing*, Cronbach (1949) similarly discussed test content as a quality criterion in achievement testing. In validity proper, he distinguished two aspects, logical and empirical. Empirical validity involved correlation with a criterion. Logical validity was based on expert judgements of what the test measured, and what was sought was ”psychological understanding of the processes that affect scores” (p. 48). This non-empirical and non-behavioral side to validation was a pre-cursor of the theoretical development in the latter half of the century.

Instead of validity, it was reliability that measurement theorists focused on. Classical test theory defined reliability as accuracy and consistency of measurement or the relationship between people’s observed scores on a fallible test and their true scores on an ideal, error-free measure of what was being tested. Reliability, like validity, was a correlation coefficient. It also defined technically the upper limit of the validity coefficient. (Angoff 1988:20, Henning 1987:90.) This was because validity involved the relationship between the “true scores” and the criterion rather than the test scores, which always included measurement error. Furthermore, the validity coefficient could only reach the upper limit of the reliability coefficient if the desired “true score” and the criterion were identical. Since this could very rarely be the case, validity would tend to be lower than the reliability coefficient – all the more so because the indicators for the *criterion* were also likely to include measurement error.

The kinds of language tests that were developed when the psychometric notions of reliability and validity were first being formed were a new breed of ”objective” tests. Spolsky (1995:33-41) says that the rise of the objectively scorable discrete-point test was a response to criticisms against the unfairness of the traditional essay examination. He cites Edgeworth’s (1888) criticism of the ”unavoidable uncertainty” of these examinations as an important motivation for the development, and mentions objectively scorable spelling tests and Thorndike’s (1903) work on the development of improved essay marking scales as important early responses in the area of language testing. A unifying theme when new tests and marking systems were developed was the desire to be fair to test takers through an improvement of the reliability of the tests.

Spolsky (1995:42-43) states that the work on the form and content of objective language tests was guided by four concerns: validity, reliability, comprehensiveness, and administrative feasibility. In practice, he reports,

administrative feasibility sometimes won over all the other concerns. Speaking and writing were considered important aspects of knowing a foreign language, but they often came to be excluded from large-scale language test batteries because it was so difficult to develop objective scoring systems for them. Hence, most objective language tests tested vocabulary, grammar, reading, and listening through multiple choice and true-false items. However, considerable effort was also spent on scales for rating written composition (Spolsky 1995:44-46). Where speaking was tested, the testing boards attempted to ensure reliability through using a board of examiners and investigating inter-rater reliability.

In sum, the early focus of validation was on the prediction of specific criteria, which later validity theory termed criterion-related validity. The content of a test was considered to be relevant proof of its validity when no obvious criterion existed for the evaluation of the test. Reliability was a prime concern, and a necessary condition for validity. The concern for the prediction of the criterion lives on in the modern version of validity theory, but it is not as central as it was earlier. Content concerns are similarly included, but they are now considered relevant for all tests, including the ones that are used to predict future performance. Reliability continues to be important for test evaluation, but it is not as clearly separable from validity in modern psychometrics (see e.g. Moss 1994:7, Wiley 1991:76 for arguments on the desirability of this development). The early focus on the usefulness of tests for specific intended uses continues to the present day.

3.3 Theoretical evolution in validity theory

From the early focus on predicting different practical outcomes or events, validity theory evolved through three or four distinct types of validity to a wide variety of validity concerns, which have lately come to be seen as aspects of a unified validity argument centred on construct validity (Anastasi 1986:1-2; Angoff 1988: 29-30; Messick 1989a:18-20; Moss, 1992:231-232; Shepard, 1993:406). Messick (1989a:18) saw this theoretical development as a reflection of a fundamental shift in the aim of validation studies. It was no longer mere quantification of the predictive power of a test. Instead, validation was aimed at sound and empirically grounded interpretation and explanation of the test scores.

The evolution in the way validity was perceived did not take place in the field of validity theory alone. At the same time, the scientific community's philosophical thinking was also changing, and the changes in the aims of validity theory reflect parallel developments in the philosophy of

science in the twentieth century. This historical development starts from positivism and moves on to relativism, instrumentalism, rationalism, critical realism, and beyond. In the historical overview that follows, philosophical dimensions will be referred to now and then, but the main emphasis is on the theoretical developments in mainstream, psychometric validity theory in the latter half of the twentieth century.

A concise overview of philosophical bases in validity inquiry is given by Messick (1989a:21-34), where the case is made that validation as an activity is not clearly tied to any single philosophical perspective but is compatible with several. Such separation of activity from philosophy has not been generally accepted in educational measurement. Maguire, Hattie and Haig (1994), for example, advocate a pure realist view and criticize Messick (1989a) for inconsistency, while Shohamy (1997) argues for the value of a critical realist approach to language testing over other philosophical stances. The fact that these different viewpoints exist, however, goes some way to prove Messick's point. Validation can be conducted under different philosophical paradigms, although the practical work that relates to validation and the purposes for which the results are used vary.

One of Messick's central points throughout his writing is that validation and values are inextricably linked. In the seminal validity chapter (Messick 1989a), he calls for the application of different systems of inquiry in validation to expose the values involved. One response to this call is Moss's (1992) introduction of a hermeneutic approach to assessment. She makes a clear distinction between the hermeneutic aim to explain a complex phenomenon, such as language skills, through the inclusion of as many theoretical concepts that are required to account for variation in performance on the one hand, and the psychometric aim to measure only those aspects which are measurable on one theoretical dimension on the other. Such philosophical discussion in language assessment is rare and has as yet not led to clear advice for test developers or validators. However, this may be one of the future trends in validity discussions, which began from scientific realism in the first half of the twentieth century.

3.3.1 Types of validity

By the middle of the twentieth century, measurement theorists were becoming concerned about the range of test development and evaluation practices which prediction-based validity allowed. They felt the need to give guidance on practice, and the division of validity into a number of separate types evolved out of this need. The way that the community chose to make

their recommendations known was the creation of professional standards. The current AERA, APA, and NCME *Standards for educational and psychological measurement* are the sixth edition; the first were published by APA in 1954 and by AERA and the predecessor of NCME in 1955 (AERA 1999:v).

Originally, theorists distinguished four types of validity: content, concurrent, predictive, and construct validity. The concurrent and predictive categories were subsequently seen to constitute subcategories of a single type, ie. criterion-related validity. The dissemination of the new orthodoxy was quick, because major textbook writers were members of professional boards, and they implemented the new terminology in their textbooks immediately. Thus, budding measurement experts began to learn about validity types in the 1950s.

The division of validity into types was a development of an earlier belief that validation, ideally a single coefficient, was different for different purposes of test use. Thus, evidence for content validity was called for when test users needed to make inferences from the test score to a content domain, for instance, how well students knew the principles of arithmetic that they had been taught. Evidence for predictive validity was required when the test was used for selection, and evidence for concurrent validity when a new test replaced an old one. The last type, construct validity, first appeared as a technical term in the 1954 *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA). At this point it was considered to be needed only when inferences were drawn from the test score to a theoretical construct that could not be defined by a content domain or a criterion variable (Moss 1992:232).

A further "type" of validity which was mentioned and briefly discussed in early textbooks was face validity or the superficial appearance of validity to non-experts who look at a test (eg. Anastasi 1954:121-122, Cronbach 1949:47). The concept was mostly dismissed as incidental because it was only based on the appearance of the test rather than logical or empirical analysis. Many authors acknowledged, however, that the appearance of a test might influence its appeal to test takers or potential users. More recently, Nevo (1985) argued that this concession makes face validity a useful concept, and showed how it might be assessed reliably. Alderson *et al.* (1995:172-173) similarly discuss the potentially influential, "acceptability" side of face validity, especially its possible motivating effect on test takers. Bachman and Palmer (1996:24) see acceptability as part of the authenticity of test tasks. Thus, while the "superficial judgement" side of test appearance is not considered to be a serious validity concern,

acceptability to stakeholders is probably an important concern, at least for test development if not for validation.

The division of validity into types unified practice, but by the 1970s, measurement theorists began to feel unhappy with the mechanical and simplified way in which commercial testing boards operationalized it. This is reflected in Guion's (1980:386) pointed summary that the three types of validity came to be seen as "something of a holy trinity, three separate ways leading to psychometric salvation." Even if the occasional testing board should pursue all three, the way in which they did so did not meet the theorists' requirements of validation, as is shown in Anastasi's (1986:2) criticism: "Thus test constructors would feel obliged to tick [the three types of validity] off in a checklist fashion. It was felt that they should be covered somehow in three properly labelled validity sections in the technical manual, regardless of the nature or purpose of the particular test. Once this tripartite coverage was accomplished, there was the relaxed feeling that validation requirements had been met."

The criticism was caused by a shift in measurement theorists' perception of what was involved in validation, especially construct validation. Loevinger had raised some concerns over the partitioning of validity evidence into four coequal categories as early as 1957. Her argument, repeated and developed by Messick (1975, 1980), was that content, concurrent, and predictive categories were possible supporting evidence for construct validity. Construct validity was the overarching term and represented "the whole of validity from a scientific point of view" (Loevinger, 1957:636) (see eg. Anastasi 1986, Cronbach 1988, Messick 1989a). This was because content validation and criterion-related validation essentially provided different types of evidence required for a comprehensive account of construct validity.

3.3.2 The rise and development of construct validation

All the theorists of educational measurement who write about the development of validity theory (e.g. Angoff 1988, Anastasi 1986, Messick 1989a, Moss 1992, Shepard 1993) ascribe the introduction of the modern concept of construct validation to Cronbach and Meehl (1955). Much more clearly than earlier writers, these authors focused on the meaning of the scores rather than prediction as the essential question in validation. Angoff (1988:26) recounts that "in construct validity … Cronbach and Meehl maintained that we examine the psychological trait, or construct, presumed to be measured by the test and we cause a continuing, research interplay to take place between the scores earned on the test and the theory underlying

the construct." In other words, the validation of a test was intimately bound with the theory of the trait or ability being measured. When construct validity was defined in this way, it could not be expressed by a single coefficient. It had to explain the meaning of the test scores. This required combining empirical evidence with theoretical statements about what the scores stood for.

Cronbach and Meehl's original (1955) concept of construct validity was expressed in the language of positivist philosophy of science dominant at the time. A construct could only be accepted if it was located in a fully specified "nomological network", which defined its relationships with other constructs and with practical observations by clear causal or statistical laws. The positivistic emphasis on the logical structure of scientific theories has since given way to instrumentalism and realism in validation research as in social sciences generally (on this development, see e.g. Messick 1989a:22-30, Cronbach 1989:158-163). The focus is more on the way in which scientific inquiry is conducted, and very similar guidelines are provided although they differ in terms of how they view truth, whether it is metaphysical and viewer-dependent or whether it exists in the world independently of any viewers.

Apart from philosophical changes, Cronbach (1975, 1986, 1988) has also argued that his and Meehl's call for fully specified relationships was both unrealistic and impractical. Less clearly specified constructs can be highly useful to explain what test scores mean. Several modern theorists of educational measurement concur with this view (e.g. Anastasi 1986, Messick 1989a, Wiley 1991). Shepard (1993:417) continues: "Nevertheless, by some other name the organizing and interpretive power of something like a nomological net is still central to the conduct of validity investigations. Perhaps it should be called a conceptual network or a validity framework." This view of the continuing importance of the conceptual network for validation is important for the present thesis. It shows how central the construct definition is in the current concept of validity.

The new formulation of construct validity as a theoretical concept was elaborated further when Campbell and Fiske (1959) presented a conceptual and empirical test which guided its operationalization. To pass the test, validators would have to assemble *"convergent evidence*, which demonstrates that a measure is related to other measures of the same construct and other variables that it should relate to on theoretical grounds, and *discriminant evidence*, which demonstrates that the measure is not unduly related to measures of other distinct constructs" (Moss 1992:233). Thus, as Angoff (1988:26) explains, an empirical design to investigate a

proposed construct would involve tests of two or more constructs tested through two or more different methods. The results might support the theory and the tests, or they might call into question the tests or test methods, the hypothesized relationship between the constructs, or the theories governing the constructs.

The logic with the call for convergent and discriminant evidence is that it offers proof both for what the test scores reflect and for what they do *not* reflect. In a further development in the theory of construct validity, Cronbach (1971) applied the same logic on a higher level of abstraction when he proposed that construct validity be guided by a search for plausible rival hypotheses. To defeat these would offer the strongest possible support for the current theory. This parallelled Popper's (1968) view concerning the development of scientific theories, as Cronbach acknowledged. Concurring with and promoting this theoretical development, Messick (1980) called the search for plausible rival hypotheses the hallmark of construct validation.

Messick (1989a:18-19) gives a detailed account of how the testing field has moved towards recognising the unitary nature of validity. He describes how the *Standards* move from an explicit division of validity into types in 1966 and 1974 to the 1985 version where validity is a unitary concept and "an ideal validation includes several types of evidence" (APA 1985:9). He points out that similar developments can be seen in successive editions of major textbooks on testing theory such as Anastasi's *Psychological Testing* and Cronbach's *Essentials of Psychological Testing*.

Messick himself was an early defender of the unified view. In 1980 (p. 1015) he argued that "construct validity is indeed the unifying concept of validity that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships." Agreeing, Cronbach (1990:152) concludes: "The three famous terms do no more than spotlight aspects of the reasoning. To emphasize this point the latest *Standards* speak not of 'content validity,' for example, but of 'content-related evidence of validity.' The end goal of validation being explanation and understanding, construct validation is of greatest long-run importance." Defined in this way, the goal of construct validation is that of science in general.

The development of validity theory in language testing has parallelled the development in educational measurement. Early textbooks such as Lado (1961) and Harris (1969) discuss internal and external validity as per Loevinger's (1957) model. In the 1970s, authors such as Davies (1977) and

Heaton (1975) considered validity in terms of four distinct types: content, predictive, concurrent, and construct. They also discussed and dismissed face validity like measurement theorists in general. Oller (1979) did not discuss validity as a theoretical concept explicitly, but his references to validity (eg. 1979:417-418) show that he considers it mainly a correlational indicator and dependent on reliability. The theoretical debate around his Unitary Competence Hypothesis (reacted to eg. by Bachman and Palmer 1982 and Upshur and Homburg 1983) was fought on empirical, quantitatively analysed evidence for construct validity. More recently, Bachman (1990) brought validity theory up for a thorough discussion and promoted the unified theory of construct validity proposed by Messick. Cumming (1996:5) summarises the development: "Rather than enumerating various types of validity ... the concept of *construct validity* has been widely agreed upon as *the* single, fundamental principle that subsumes various other aspects of validation ... relegating their status to research strategies or categories of empirical evidence by which construct validity might be assessed or asserted." When Bachman and Palmer (1996:21) list the essential properties of language tests, they mention and define *construct* validity: "Construct validity pertains to the meaningfulness and appropriateness of the *interpretations* that we make on the basis of test scores."

To summarise, the shift from prediction of specific criteria to explanation of the meaning of test scores as the aim of validation raised construct validity to a central position in validation inquiry. At the intermediate stage, theorists identified three or four types of validity of which construct validity was one, but several theorists made compelling cases that the other "types" of validity could not sustain a validity argument alone and were better seen as aspects of a unified theory of construct validity.

3.4 The current concept of construct validity

In what follows, I will summarise the core of current validity theory. I will explain why construct validity is considered to be the central concern. I will also cover theorists' views of threats to construct validity. Next, I will discuss the current emphasis on validity theory on the validation process rather than validity as a property. This makes validation tangible, because the discussion concentrates on the activities that validators do. Furthermore, I will give a brief overview of the main areas of current debate in validity theory: how construct validity can or should be organised into component

parts, whether social consequences of test use are a relevant concern for construct validity, and whether performance assessments, which are becoming more and more common in educational measurement, require their own validity criteria. In all of the discussion below, I will give special emphasis to Samuel Messick's views of validity and validation, because he has been one of the most influential thinkers and writers on validity in educational measurement in the latter half of the twentieth century.

3.4.1 The centrality of construct validity

Measurement theorists consider construct validity to be the main validity concern because the object of validation is seen to be the interpretations of test scores, and because the current view is that the interpretations necessarily involve constructs (eg. Cronbach 1980 and Messick 1975, 1980 argue for this view). Many researchers consequently use the terms validity and construct validity interchangeably.

As discussed in the introduction, constructs are theoretical concepts that tests are considered to implement. In foreign language tests, relevant constructs might include language ability, reading, and ability to comprehend the main idea or ability to draw inferences from reading material. They might also include more "functional" constructs such as reading for information or reading to summarize. Tests implement the constructs through the tasks and assessment criteria, and they produce scores as indicators of the construct.

In current validity theory, validation is focused on scores rather than tests. This is because the interpretations or inferences in the use of assessments are drawn from the scores, not the instruments (Messick 1989a:14). The scores reflect the properties not only of the assessment instrument but also "of the *persons* responding and the *context* of measurement" (Messick 1989a:14). All the systematic influences that can affect scores should be investigated in a validation exercise. Thus the scope of validation inquiry is quite broad. In addition to the test scores themselves, the test, the testing procedures, the context in which the test is implemented, and the processes that the test takers and assessors go through during the testing process must be investigated to explain the meaning of the scores. From the perspective of the testing board, all the objects of investigation are nevertheless related to their test and its implementation and use, which offers a concrete basis for the studies.

Scores and constructs are related, but they are on different levels of abstraction in the network of testing and score interpretation. The score, like the test and testing situation from which it resulted, is one of an extensible

set of concrete operationalizations of the construct. The construct is abstract and more generalizable than the score or the test. The scores from a test reflect not just the construct of interest but also other factors. They can be influenced by other constructs instead of, or in addition to, the one that the users are interested in. For instance, they can be influenced by the format of the test or the degree of time pressure that the participants were under when taking the test.

3.4.2 Threats to validity

Cook and Campbell (1979) distinguish two kinds of threats to validity: construct underrepresentation and construct-irrelevant variance. This is an elegant way of formulating the main concerns. In construct underrepresentation, the threat is that the test is too narrow and the score does not reflect enough aspects of the construct of interest. In construct-irrelevant variance, the threat is that there are other influences which are independent of the focal construct but which are consistently affecting the scores, while the scores are interpreted only in terms of the focal construct.

Construct underrepresentation and construct-irrelevant variance appear opposite in terms of the content covered in the test, but it is important to note that they do not cancel each other out. Messick (1995:742) explains that "both threats are operative in all assessments. Hence a primary validation concern is the extent to which the same assessment might underrepresent the focal construct while simultaneously contaminating the scores with construct-irrelevant variance."

Construct underrepresentation and construct-irrelevant variance are serious threats to validity because they can lead to distorted interpretations of test scores, which in turn may cause adverse consequences on some individuals or groups taking the test. However, all adverse consequences are not necessarily the result of problems with test validity; it can also be that the low scores correctly describe the particular groups or individuals who are negatively affected. Bias, or adverse consequences for particular subgroups, must be guarded against. However, according to the theorists, adverse consequences are validity concerns only if they can be traced back to irrelevant sources of test and criterion variance, that is, construct underrepresentation or construct-irrelevant variance (Messick 1989b:11, 1995:748).

3.4.3 Concentration on the validation process

Much of current validity theory concentrates on validation as a process rather than validity as a property. Anastasi (1982, 1986), for instance,

discusses types of *validation procedures* and the inclusiveness of construct valid*ation*, Cronbach (1988, 1990) describes and discusses validity *inquiry* and validity *argument.* For Messick (1989a:19), this reflects the increasing favour of the unitary concept of construct validity because by choosing this terminology, the authors make it clear that unitary validity has to be supported by different kinds of validity-related activities. The 1985 *Standards* define validation as "the process of accumulating evidence to support [score-based] inferences" (p. 9), while the 1999 *Standards* states that "the process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations" (p. 9).

Validation work begins when the idea for a test is born and ends when the scores are no longer used. According to Anastasi (1986: 12-13), "there is a growing recognition that validation extends across the entire test construction process; it encompasses multiple procedures employed sequentially at appropriate stages. Validity is built into a test at the time of initial construct definition and the formulation of item-writing specifications; the hypotheses that guide the early developmental stages are tested sequentially through internal and external statistical analyses of empirical data."

Cronbach (1990:183-184) distinguishes between strong and weak construct validation. The weak approach involves building a validity case out of any and all available evidence in support of a desired score interpretation. The strong approach is more structured and more clearly driven by theory. It calls for expressing the theoretical underpinnings of the scores as clearly as possible, supporting the theories and hypotheses with empirical evidence, trying to resolve crucial uncertainties, and defending the proposed interpretation(s) against plausible counter-interpretations. The weak approach is not without merit, Cronbach says, but it lacks the purposefulness of the strong one. The heart of the strong approach is the recognition of plausible rival hypotheses.

Furthermore, Cronbach invites educational measurement professionals to think about validity in terms of evaluative argument rather than in terms of validation research (1988:4, 1990:185-189). "Argument" implies a human protagonist making a case as persuasive as possible in front of a critical audience. The source and motivation for Cronbach's thinking is litigation. Increasing numbers of cases are brought into court, especially in the United States, against tests which have been used to make social decisions. On trial in such cases is a particular interpretation in the context of a decision making process. Cronbach makes a distinction between the user's interpretation in a particular case and the work that a test

developer has to put in to prepare for possible litigation. Argument serves as a rationale for the test developer's work because they have to check that cases can be made on the basis of the information that they provide. The test developer has to "lay out evidence and argument that will help the entire profession make sense of scores from the test. Users will want to know about the processes required for successful test performance, about the relation of this score to traits that are better understood, about background factors associated with good and poor scores, and so on. Such information helps them to recognize what alternative interpretations of scores are plausible wherever they use the test" (Cronbach 1990:189).

Current theory stresses the importance of *evidence* for validity cases. Theoretical rationales for certain meanings and interpretations should be presented, but these should be supported by different kinds of empirical evidence. The evidence can be qualitative or quantitative, and it should represent several different perspectives on the meaning of the scores. The more lines of evidence that support an interpretation of the scores, the better. If the lines of evidence conflict, ie. if some evidence supports a desired interpretation and some does not seem to be relevant or seems to conflict, this is highly useful for construct validation as well. Such situations lead the validators to reconsider the aim of the test and the meaning of the scores, and this is exactly what construct validation should do.

The essence of the validation process is nicely summarised by Cronbach:

> Construct validation is a fluid, creative process. The test constructor or any subsequent investigator works to develop an interpretation, persuade others of its soundness, and revise it as inadequacies are recognized. Self-criticism and criticism from persons preferring other interpretations play an important role. The interpretation has scientific aspects, but it often embodies policies and suggests practical actions. This complexity means that validation cannot be reduced to rules, and that no interpretation is the final word, established for all time.
> Cronbach 1990:197

This description puts the construct at the centre of the world of test development, validation, and use. It implies links between testing, construct validation, and the society that uses the scores, and places considerable requirements on the validators and the society around them. The validators must be able to retrieve the construct rationale from existing documentation, know what would constitute "inadequacy" in interpretation, acquire criticisms, revise the interpretation if required, and have enough status to influence the testing board and the society using the scores to establish the required changes. Cronbach refers to the evident complexity of the activity. To make it possible for practitioners to try to accomplish this, they must

use the construct definition as the organising principle for all the activities related to the development and validation of the test. Some form of a construct definition should be written down as soon as possible, and this definition should be revised and extended in the course of the work. Most importantly, all the activities in test development, validation, and use should always be linked back to the current formulation of the construct.

3.4.4 The complexity of unified validity

Since the division of validity into three types with quantified indicators was abandoned and construct validity began to be seen as the main concern, validity has become a complex concept. One of the fathers of the current concept, Samuel Messick, would argue that this is a well motivated complexity. He sees tests as instruments which are used in society, and his point is that the meanings of test scores cannot and should not be investigated without reference to the way they are going to be used. Instead, investigations of validity should always entail inquiry into the values and consequences involved in the interpretation and use of test scores. He has promoted a unified view of validity throughout his writings (eg. 1975, 1980, 1982, 1984, 1989a, 1989b, 1994, 1995), but to make it more comprehensible, he has also proposed a new model for the concept.

Messick calls his faceted conception of unified validity the progressive matrix (see Table 1). He distinguishes two main facets in testing: the source of justification for the testing, which can be either evidence for score meaning or consequences of score use, and function or outcome of testing, which can be either test interpretation or test use. The heart of this formulation is score meaning, but its four conceptual categories express Messick's three main theses about the nature of validity: (1) that values form an integral part of score meaning, (2) that both the theoretical meaning arising from the measure and the applied meaning which is connected to particular contexts of test use need to be considered in construct validity, and (3) that consequences of test use form an essential aspect of score meaning.

Table 1. Facets of Validity as a Progressive Matrix (Messick 1989b:10)

	Test Interpretation	**Test Use**
Evidential Basis	Construct Validity (CV)	CV + Relevance/Utility (R/U)
Consequential Basis	CV + Value Implications (VI)	CV + R/U + VI + Social Consequences

The first cell, the evidential basis for test interpretation, calls for evidence for score meaning, which is the core meaning of construct validity. The kinds of evidence that belong here are content relevance and representativeness, theoretical comprehensiveness of representation, correspondence between theoretical structure and scoring structure, relationships between items within the test, and relationships between scores or sub-scores and other measures (Messick 1989a:34-57). The second cell, the evidential basis for test use, requires additional evidence for the relevance and utility of the scores for a particular applied purpose. Construct validity belongs to this cell because the relevance and utility of score meaning for any purpose are dependent on the evidential meaning of the score. The consequential basis of test interpretation calls for considerations of the value implications of score interpretation, including the values of construct labels, theories, and ideological bases of the test. The consequential basis of test use requires the evaluation of the potential and actual consequences of score use.

Messick calls the matrix progressive, because construct validity appears in each of its cells. In the previous version of the matrix, construct validity had only appeared in the first cell, although in explaining the figure, Messick (e.g. 1980:1019-1023) stressed that the other cells illustrated specific aspects of score meaning. By including construct validity in all the cells in 1989b, Messick clarified a disjunction between the figure and the explanation. The inclusion of construct validity in all the four cells emphasizes the centrality of construct meaning in Messick's conception of validity.

While agreeing that the social dimensions that Messick introduces to validity are important, Shepard (1993, 1997) criticizes the matrix formulation because it is conceptually difficult to understand. Construct validity appears in every cell, yet the whole matrix also depicts construct validity. Moreover, she argues that the progressive nature of the matrix allows investigators to begin with "simple" construct validity concerns in the first cell, and they may never get to the fourth cell where consequences of measurement use are addressed. She says that this is unfortunate because it is not at all what Messick intended, but this is the way the matrix is sometimes used. Moss (1995:7) similarly agrees with the importance of the social meaning of scores, but suggests that the progressive matrix cannot replace the traditional categories of content, criterion, and construct-related evidence because it does not distinguish categories within the concept of construct

validity. Rather, it locates construct validity in a larger notion of validity which includes values and consequences.

Chapelle (1994) applied Messick's concept of construct validity to evaluate validity when c-tests are used in research on second language (L2) vocabulary. Her evaluation covered all the cells of Messick's matrix, ie. the four concerns of construct validity, relevance and utility, value implications, and social consequences. Following Messick's theory, she began the investigation by defining the construct of interest, vocabulary ability. Throughout her analysis, she referred to this construct definition, using it as a criterion in the evaluation. She discussed the first cell, construct validity, through six types of evidence and analyses: content evidence, item analysis, task analysis, internal test structure, correlational research, and experimental research identifying performance differences under different theoretical conditions (Chapelle 1994:168-178). Although Chapelle did not investigate a single test but a test method, and the immediate context of reference was second language acquisition (SLA) theory rather than the use of examinations for decision-making purposes in social life, her faithful application of Messick's concept of validity showed that the theory can be understood and operationalized.

Two observations from Chapelle's study are particularly relevant for the present thesis: firstly that the guiding force in her study was the detailed construct definition, and secondly that the article actually defined a research programme for a thorough evaluation of the validity of using c-tests in research on L2 vocabulary. The programme is similarly based on the construct definition. The implication from the first observation is that a detailed construct definition can provide an elegant design principle for a coherent study. The implication from the second observation is that a construct-driven rationale can lead to a very broad research agenda. This is by no means a demerit; however, it is too big a challenge for an individual test development board. Lines will thus need to be drawn between what is immediately relevant for testing boards and what is part of a broader discussion.

3.4.5 Social consequences as a concern for test use

Messick is a strong advocate for the inclusion of social consequences as an integral validity concern, and other researchers who write about this always refer to his treatment of the topic. Messick recognises that the arena of social consequences is broad, and he is careful in drawing a line between validation and social policy. In the final analysis, the key concern for validation is with adverse social consequences to individuals and groups,

and concerning these, the validator's responsibility ends with the quality of the instrument:

> A major concern in practice is to distinguish adverse social consequences that stem from valid descriptions of individual and group differences from adverse consequences that derive from sources of test invalidity ... The latter adverse consequences of test invalidity present measurement problems that need to be investigated in the validation process, whereas the former consequences of valid assessment represent problems of social policy. Messick (1995a:744)

Where the limit lies in practice and how far the responsibility of the test developer extends are perhaps less clear-cut issues in practice than in this abstract distinction. This is because both realised and potential consequences of test use should be investigated and negative consequences avoided.

In the case of realised adverse consequences, scores have already been used, and someone makes a claim of injustice. Messick's (1989a:85) example for such a case relates to gender or ethnic differences in the way the scores are distributed in a university entrance test. The question is how the differences can be explained: whether gender or ethnicity actually influences the test scores, or whether the desired characteristics are measured in the entrance test and those who do not pass the test fail because they lack the essential characteristics which are required for university study. The overall aim is to establish whether injustice has happened, and if so, to identify its source. In such a case, the test developers or validators hope to prove that the source is not test invalidity.

Potential social consequences are equally related to the test and its background, but as the concept implies, there are no concrete data for the consequences yet. Messick (1989:85) argues that this should nevertheless be considered an essential part of validation because consequences, both intended and unintended, contribute significantly to the meaning of the score when scores are used. Consequences should be considered before scores are used in reality, regardless of the difficulties involved in prediction, because this might reveal the kinds of evidence that will be needed for the monitoring of actual consequences once the test is used, and because serious consideration of potential consequences might help test developers and users to "capitalize on positive effects [of score use] and to ameliorate or forestall negative ones" (Messick 1995a:744).

Messick (1989:85-86) makes two suggestions as to how to identify potential consequences: deriving hypotheses from the construct meaning, and considering the consequences of *not* using the test but doing something else instead. The "something else" might be an alternative assessment mode

such as on-the-job observation, or the decision not to assess at all. A factory might decide to train all its workers instead of testing them and promoting skilled ones (Messick 1989:86). The alternative solutions to using tests also have consequences, in terms of costs as well as values, as Messick points out. It may be difficult to change existing values, but it is illuminating to analyse them in the light of alternative solutions.

As emphatic as Messick is about the importance of consequences being considered, he is equally emphatic that consequences are only one of the concerns which need attention in validation. He states explicitly that "this form of evidence should not be viewed in isolation as a separate type of validity, say, of *consequential validity*" (Messick 1995b:7). Rather, it is one important aspect of construct validity, along with other, equally important aspects. I will present Messick's classification of the six aspects of construct validity later in this chapter (for a more detailed discussion, see eg. Messick 1994:22; 1995a:744-746; 1996:248-253).

The question of whether social consequences should be considered integral to validity is an open issue, however, and this has caused a heated debate among the measurement community at the end of the twentieth century. Theorists agree that validity is the most important consideration when tests are evaluated, but they do not agree on the scope of the concept.

3.4.6 Social consequences as integral concerns for validity

None of the theorists in educational measurement contests the idea that social consequences are an important concern when tests are used to make decisions in society. The question is whether social consequences are a separate concern of test use or an integral part of validity. Those who are against their inclusion in validity (Maguire, Hattie, and Haig 1994; Mehrens 1997; Popham 1997) argue that concern for consequences unnecessarily clutters the concept of validity. Those who favour the inclusion (eg. Linn 1993, 1997; Messick 1975, 1980, 1989, 1995a; Moss 1995; Shepard 1993, 1997) argue that social consequences should be included in validity because they are so important to the evaluation of tests. In the words of Linn (1997:16), "Removing considerations of consequences from the domain of validity … would relegate them to lower priority. Validity is, after all, … the most important consideration in test evaluation."

The debate around the status of social consequences as concerns for validity is highly polarised. Those who argue for the inclusion are strongly *for* this development, those who are against it are clearly against. Proponents of the broader, inclusive view stress that test use has always been considered important in validation. Social consequences are integral

validity concerns because most tests are used to make decisions in the social world, and the hypotheses supporting this use belong to the set of inferences which must be supported in a validity argument. The advocates of the narrower view consider the investigation of the use and misuse of tests socially important but a matter of ethics and social policy rather than validation.

Theorists who would like to define the limits of validity narrowly tend to stress that validity is focused on the *accuracy* of test-based inferences (e.g. Popham 1997, Mehrens 1997). The measurement community seems to be fairly well agreed, however, that validity cannot be limited to accuracy only, it also involves decisions on appropriateness and usefulness. These aspects are key to Messick's definition of validity, for instance. Cronbach (1988:4) makes a strong case for validation being a form of evaluation, which involves determination of "truth" but also arguments and judgements about "worth". Accuracy contributes to usefulness but does not guarantee it, because a measure can be reliable and accurate but not appropriate for a particular purpose (Cronbach 1988:5). But considerations of worth do not necessarily entail the inclusion of consequences of test use into validity.

Cronbach (1998:27) reports that the new edition of the *Standards* is going to take the narrow view. "The forthcoming edition of the Test Standards, assuming it is not changed from here on radically, handles [the consequences of using tests] by just saying flatly that the Standards are going to stop with the scientific interpretation of the testing and not deal with consequences. Consequences are important but not part of the validity of using a selection test routinely, mechanically, without judgment. As for the consequences it has for eliminating certain populations from the group served, important, but not part of test validity." Cronbach goes on to explain that this means that the validator's task is to list the choices available to the users. The community of users must bear the responsibility for the consequences.

Cronbach's statement is largely borne out in the new *Standards*, although the position is not quite as clear as he makes it. Differential consequences of test use are a validity matter if they "can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components" but not otherwise (AERA 1999:16). The *Standards* further points out that if claims are made that a test introduces benefits for the users, such as prevention of unqualified individuals from entering a profession or the improvement of student motivation, then an important part of the validation programme for such a test is gathering evidence that the beneficial consequences are actually realised (pp. 16-17).

3.4.7 Validity in performance assessment

A debate which is related to the heightened attention to the consequences of test use is that of the value and validity of performance assessments. In the late 1980s and early 1990s, researchers like Frederiksen and Collins (1989) and Linn, Baker, and Dunbar (1991) proposed specified validity criteria to performance based assessments. They felt that the generic criteria for evaluating the quality of tests placed too much emphasis on reliability, which, defined as quantification of consistency among independent observations, requires a significant level of standardization. Such values favoured selected response tests over performance based tests, they argued, without regard to the educational effects that the use of these tests had.

To change the balance towards increased consideration of consequences of measurement use, Frederiksen and Collins (1989:27) proposed the notion of ”systemic validity”. This referred to the effect that the nature of the test had on the educational system of which it formed a part. Linn et al. (1991:17-21) presented a long list of criteria which should be considered in evaluating a test: consequences for teaching, fairness to test takers, transfer and generalizability of the skills required by the test taking to the skills of interest, cognitive complexity of tasks, the quality and coverage of the task materials, meaningfulness of tasks to test takers and their teachers, and the cost and efficiency of the testing activities. In this view, testing is seen as an activity, and the aim is to relate it as positively as possible with other activities in the classroom. Similar cases about the effects and values of testing have been raised in language testing under the general notions of authenticity, directness, washback, and impact, as was discussed in the previous chapter.

Performance assessment is often promoted as a way of increasing the authenticity and directness of assessment procedures. Messick (1994) interprets these calls as arguments that promote the generic validity criterion of construct representation. He sees the authenticity issue as a call for minimal construct underrepresentation, and directness as a call for minimal construct-irrelevant variance (Messick 1994:14). Frederiksen and Collins (1989) and Linn, Baker and Dunbar (1991) seek to promote performance assessment because they are worried that technical validity criteria are used narrowly and unthinkingly to exclude sensible types of assessment which are highly suitable for classroom contexts. Messick (1994) argues that these technical validity criteria now include authenticity and directness in the form of construct representation. He continues that what is needed is evidence

and arguments to support the case of performance assessments on the validity arena. However, from Messick's point of view, this requires that performance assessment is seen in "competency- or construct-centered" rather than "task-centered" terms (p. 14). If the assessment focuses on individual task performances rather than abilities, task-constructs abound and cannot provide useful, generalizable meanings for scores given to persons about their abilities, knowledge, or processes. Such person-centred measurement is central in much of current educational measurement including language testing, while sociocultural definitions of constructs which are more tied to contexts and tasks are more marginal. This may be because of the strong tradition in educational systems to give scores to individuals. Task-based assessment might be useful for the evaluation of educational systems, but such assessments are not nearly as common in current educational systems as the assessment of individuals. The prevalence of individual-based assessments may also be related to the current lack of models for ways in which a range of task-based or performance-based assessments might be combined to provide information about individuals in interaction with others and with their contexts. Some such models or guidelines would be required to develop measurement models for the new kind of constructs or construct groups. Current psychometric models can deal with individually based measurement because the subject on which generalizations should be made is well defined, while task-based assessment with its open range of constructs is a challenge in this respect.

3.5 Approaches to validation

The theoretical literature on validity gives fairly concrete and detailed advice on the *methods* of validity inquiry. The list contains most if not all research methods used in social science, which is not surprising because construct validation is modelled after scientific research in general. The problem for the test developer, validator, or user is not really one of choosing the research methods. Rather, the problem is forming an overall understanding of the broad concept of validity and deciding how to implement or evaluate a concrete validity case.

Theorists increasingly favour Messick's matrix formulation and Cronbach's strong program of construct validation as guiding frameworks for validity inquiry. Messick's matrix emphasizes the construct rationale, multiple sources of evidence, and attention to the consequences of testing. Cronbach's strong program is grounded in an explicit conceptual

framework and produces “an integrative argument that justifies (and refutes challenges to) the proposed meaning of the test score” (Moss 1995:7). The two proposals cohere very well, and the theorists refer to each other’s work. The challenge for those working in educational measurement is finding reasonable ways to implement the complex programs in validation practice.

The best known theoretical framework for the componentialization of the validation process is the traditional content-construct-criterion division. The 1985 *Standards* (APA 1985) presents validity data under these categories, as do most measurement textbooks. The traditional division has been criticised on many accounts, however (see eg. Anastasi 1986, Cronbach 1988, Messick 1989a, Shepard 1993). Firstly, its categories are not logically distinct or of equal importance since construct validation subsumes the other categories. Secondly, it does not help structure a validity argument. Cronbach (1989:155), for example, criticizes validity claims in test manuals which ”rake together miscellaneous correlations” instead of reporting ”incisive checks into rival hypotheses, followed by an integrative argument”. Thirdly, the traditional categories do not help answer the question ’How much evidence is enough?’. Alternative models are thus required.

In fact, several different kinds of models might be useful for test-related validation, and several models are also proposed in the educational measurement literature. One approach to the modelling of validation is to see it as a temporal, staged process. This approach is not very common, but it is particularly useful for testing boards which are developing new tests. The approach is taken by Cole and Moss (1989), who provide a framework for the process of gathering validity data during test development.

In addition to this, test developers need to use a set of principles for organising and prioritising their validation activities. Messick (1994, 1995a, 1996) proposes a set of six aspects of validity to guide validation activities. Chapelle (1994, 1999) implements Messick’s proposal, but reformulates the six aspects as “approaches to validity evidence” (Chapelle 1999:260). The current version of the *Standards for educational and psychological measurement* uses a very similar concept with five “sources of validity evidence” (AERA 1999:11).

Yet another model for the presentation of validity arguments is outlined by Kane (1992). His approach works from the proposed score interpretation “backwards”, and it relates the proposed interpretation closely to the context where the scores are used. This is useful for the construction

of validity arguments once concrete situations of score interpretation can be identified. Taken together, these proposals offer validators: guidance for the implementation of the validation process from bottom up; monitoring the quality of their activities while the process is going on; and presenting validity arguments when justification is needed for specific proposed interpretations.

3.5.1 Framework for accumulating validity data

Cole and Moss's (1989) proposal of how to organise the validation process is based on the time line of the development of an educational test. The process "typically begins with specification of a proposed use and definition of one or more constructs relevant to that use. A test is then developed with some particular content and with questions and answers in some particular format. That test must be administered and scored. How test takers respond to that test is reflected in the internal structure of items or parts of the test and in external relations of scores to other variables" (p. 205). From this process, Cole and Moss identify key objects and activities to be investigated in the validation process. These are: contextualised construct definition; content and format of the test; administration and scoring of the test; internal test structure; and external test relationships (Cole and Moss 1989:205). The investigation of each of these should be guided by hypotheses about what the test measures. The authors stress that the validators should consider both logical and empirical evidence and convergent and discriminant evidence.

Cole and Moss's (1989) proposal is interesting because it takes a single test and the interpretations available from it as its starting point. It uses the time line of the test development process as an anchoring device, thus making it possible to follow Anastasi's (1986) suggestion of that validation should begin as soon as test development begins. When a framework for the accumulation of validity data is constructed in a time-bound manner, it helps the test developers conduct and structure the validation process alongside the test development process. This, in turn, helps guarantee that when the test developers need to make an actual validity argument, the data from the validation process is available for it.

Anastasi (1986) argues for a very similar view of the validation process and similarly relates it to the test development process. According to her, an ideal validation process for a psychological test "begins with the formulation of detailed trait or construct definitions, derived from psychological theory, proper research, or systematic observation and analyses of the relevant behavioral domain. Test items are then prepared to

fit the construct definitions. Empirical item analyses follow, with the selection of the most effective (ie. valid) items from the initial item pools. Other appropriate internal analyses may then be carried out, including factor analyses of item clusters or subtests. The final stage includes validation and cross-validation of various scores and interpretive combinations of scores through statistical analyses against external, real-life criteria" (Anastasi 1986:3). Anastasi's ideal process thus initially uses the construct definition and the test as the primary sources of evidence, and when scores are available, these become the major object of interest.

3.5.2 Components of validity inquiry

Messick (1994, 1995a) proposes a set of six general validity criteria, which could be used to organise validity arguments and to judge their completeness. The categories are content, substantive, structural, generalizability, external, and consequential aspects of construct validity. The six categories are less overlapping than the traditional ones, but so far they have the disadvantage of being "new". The names "substantive" and "structural" are therefore opaque as terms. Chapelle's (1999:260-262) recasting of the categories as "approaches to validity evidence," as explained below, offers slightly more transparent terminology. The current *Standards* (AERA et al. 1999) offers a closely related set of five "sources of validity evidence", also presented below. These three sources include consequences as one of the concerns in validation, thus making validation inquiry broader than some theorists would prefer. However, since all agree that consequences must be considered at some point in evaluating tests and their use, the inclusion simply means that evaluation of test use is encompassed in this view of validation.

Messick's *content aspect* of validity refers to the degree to which the assessment tasks can be proved relevant indicators of the construct invoked in test interpretation, and representative of that construct. Judging the relevance of the assessment tasks involves specifying "the knowledge, skills, attitudes, motives, and other attributes to be revealed by the assessment tasks" as well as specifying the boundaries of the target of assessment, or what is not going to be assessed (Messick 1995a:745). Judging the representativeness of the assessment tasks involves ensuring that all important parts of the construct are covered by the test. Chapelle (1999:260) terms this approach *content analysis*, and the 1999 *Standards* (p. 11) calls it *evidence based on test content.* The documentation of expert professional judgement provides evidence for this aspect. All the authors

refer to the test's specification as a source for categories on which expert judgement is needed.

The substantive aspect refers to the relationship between the measure and its theoretical underpinnings in process terms. Messick (1995a:745) stresses that this comprises both building theoretical models of the processing involved and "empirical evidence that the ostensibly sampled processes are actually engaged by respondents in task performance". He suggests a range of sources for such evidence, including verbal protocols, eye movement records, correlation patterns among part scores, consistencies in response times, and mathematical or computer modelling of task processes. Looking at the analyses proposed, Chapelle (1999:261) calls this aspect *empirical item or task analysis*. She stresses that both quantitative and qualitative strategies are required to inquire into the skills actually assessed in a test. The *Standards*, from a slightly different perspective, call this aspect *evidence based on response processes* (1999:12). They consider this aspect to cover theoretical and empirical analyses of response processes. The results, according to the *Standards*, can clarify "differences in meaning or interpretation of test scores across relevant subgroups of examinees" (p. 12).

The structural aspect of validity focuses on the relationship between the theoretical structure of the construct and the scoring system used in the test. This concerns both how individual items are scored and how the final scores are assembled. Take, for example, a ten-item measure of attitudes towards the speakers of a language that is foreign to the respondent. In this questionnaire, each question is scored on a 5-point Likert scale and the final score is the average of the ten responses. Under the structural aspect of validity, the validators should consider whether using a 5-point scale for individual situations is coherent with their theoretical understanding of what these kinds of attitudes are like and how they might be measured. They should also consider whether averaging across ten responses accords with their theory of how case reactions relate to generic attitudes. Perhaps the response scale should have seven points or perhaps four. Perhaps means do not express the construct of ethnic attitudes adequately, perhaps range should be reported as well. The structural aspect of construct validity calls for evidence that the decisions about the scoring mechanism have been taken advisedly. The name for this aspect is derived from the comparison between the structure of construct theory and the structure of the scoring system. This is such an abstract notion, however, that the name risks being opaque. The *Standards* calls this aspect *evidence based on internal structure*, while Chapelle (1999:261) refers to this group of analysis

techniques as *dimensionality analysis*. None of the terms are perfectly clear without further explanation, but the *Standards* approach might be the most familiar to practitioners because it repeats terminology from the 1950s. The difference in the current term is the comparison to the structure implied by the construct.

The generalizability aspect concerns the "generalizability of score inferences across tasks and contexts" (Messick 1995:746) and particularly focuses on the limits beyond which the score inferences cannot be generalized. The contexts across which the generalizability of a score should be assessed empirically cover eg. different testing occasions and different assessors as well as different tasks. Messick identifies this aspect of validity with traditional reliability concerns. He reconceptualizes the problem of the traditionally recognized tension between reliability and validity (1995a:746) as a tension between "the valid description of the specifics of a complex task and the power of construct interpretation." In other words, test developers might like the test and the score to reflect the complexity of task performance, but they would also like the performance to generalise to a reasonable range of tasks which were *not* tested, and say what the scores mean in terms of general concepts such as the individual's intelligence or his/her ability to deal with customers in the language tested. Messick (1995a) does not propose methods for the generalizability aspect of validity, but he does mention traditional reliability and generalisation across tasks, occasions, and raters. Chapelle (1999:262) terms these investigations as studies of *differences* in test performance and proposes that generalizability studies and bias investigations belong to this aspect of validity. The *Standards* does not treat this group of analyses as a separable aspect of validity but discusses it together with the next group.

The external aspect of validity concerns the relationship between the scores from the measure and other phenomena, including other measures. Specifically, the external aspect is concerned with the estimation of the fit between expectations of relationships between scores and other phenomena formed on the basis of the theory behind the test and the actual empirical relationships that are found in practice. If the theory behind the test is sound and comprehensive, "the constructs represented in the assessment should rationally account for the external patterns of correlations" (Messick 1995a:746). This, of course, only holds if both the measure and the phenomena to which it is related are conceptualised in construct terms, if the expectations of the relationships are formed explicitly, and if the expectations take into account the overlap between the constructs. In addition to comprehensive theory, this calls for skilful operationalization of

the "other phenomena" to which the test should be related. Messick stresses the importance of both convergent and discriminant correlation patterns when evidence for the external aspect of validity is sought. Chapelle (1999:262) considers these studies investigations of *relationships* of test scores with other tests and behaviours, while the *Standards* (1999:13) discusses them as *evidence based on relations to other variables*. Because of the broader scope that the *standards* allocates to this category of validity evidence, their characterisation of the techniques involved is also broader. It includes convergent and discriminant evidence, test-criterion relationships, and validity generalization (pp. 14-15).

The consequential aspect of validity refers to the "evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use" (Messick 1995a:746). Through this aspect, Messick encourages test developers to both think of the consequences at all stages of test development and collect evidence for the positive and negative consequences. Chapelle considers these studies *arguments based upon testing consequences*. She points out that the arguments involve the value implications of score interpretations as well as social consequences. The *Standards* terms the aspect *evidence based on consequences of testing* and stresses the investigation of construct underrepresentation and construct-irrelevant variance as a potential source of bias in test use. Furthermore, they point out that investigation of a test's claims of positive impact belong to this category of validity studies (AERA 1999:16).

Messick concludes his presentation of the six aspects of validity by stressing the need for the validation argument to be comprehensive and either cover all the six bases or present a rationale for why some of the bases do not need to be covered. He concurs with Cronbach's (1988) view of validation as evaluation argument and examines Shepard's (1993) validity cases in terms of how well they cover his six aspects of validity. Messick argues (1995a:747) that the six aspects are consonant with, but more basic than, Kane's (1992) categories of interpretive argument (see below for a presentation of these). Messick argues that this is because score interpretation and Kane's interpretive arguments invoke or assume his six validity criteria. Moss (1995) views such comparison as an argument on the principle of organisation that measurement theorists could or should use in categorising validity. Kane's proposal is organised in terms of assumptions necessary to justify a proposed use, Messick's by aspects or targets of validity inquiry.

A potential problem with Messick's six aspects of construct validity is that they define a very large area of inquiry for test developers. Raising

this concern, Moss (1995:7) points out that there is a need to draw clear distinctions between the responsibilities of the developers of an individual measurement instrument to justify a score interpretation and the responsibilities of the scientific community at large to enhance theory and practice in the long run (1995:7). Although validation is a long and challenging process, the challenge for individual testing boards must be reasonable. She continues that making the challenge meetable could encourage both rigorous conceptualisation and reporting of validity research and "explicit attention to existing theory and research on similar assessments." She thus implies that testing boards might make the validation challenge manageable through more reference to each others' work. In this vein, the new *Standards* states that "use of existing evidence from similar tests and contexts can enhance the quality of the validity argument, especially when current data are limited" (AERA et al. 1999:11).

3.5.3 Building a validity case

Kane (1992) regards validation as a kind of practical argument. He derives his ideas from Cronbach's (1988, 1989) proposals around validity arguments, Toulmin, Rieke and Janik's (1979) ideas about practical reasoning, and House's (1980) thinking of evaluation argument. Kane explains that practical argument is a good model for validation because unlike traditional logical or mathematical arguments, practical arguments cannot be proven or verified in any absolute sense. The best that can be done is to show that the practical argument is highly plausible, which is just what happens when a validator argues for a test score interpretation. (Kane 1992:527.)

The value of Kane's (1992) proposal is the logic that he offers to validators for a specific validity argument and the investigations related to it. He suggests that validators begin by listing the statements and decisions to be based on the test scores. Next, they should specify the inferences leading from the test scores to these statements and decisions and identify potential competing interpretations. Finally, they should assemble evidence supporting the main argument and refuting the potential counterarguments (Kane 1992:527). They can evaluate their case by three general criteria for the evaluation of practical arguments: clarity of argument, coherence of argument, and plausibility of assumptions. When they plan additional studies to strengthen their case, they must focus on the weakest inferences first as these are the most vulnerable to criticism and counterarguments (Kane 1992:528). Shepard (1993:432) reasserts the importance of this recommendation. Once the separate assumptions have been laid out and the

evidence organised accordingly, Kane's design makes it clear that if some of the assumptions are completely unsupported, no amount of support for the other assumptions strengthens the case for the proposed test use. The argument stands or falls by its weakest link. This is a valuable if challenging reminder for test developers.

Kane (1992:531-532) gives a constructed example of a validity argument. This involves a placement test which is used to place students into a calculus course or a remedial algebra course. He lists seven main assumptions which the proposed test interpretation builds on. The assumptions concern the notion that algebraic skills really are a prerequisite for the calculus course, the test as a measure of these skills, appropriate placement for students with low placement scores, and appropriate placement for students with high placement scores. The array of assumptions makes it clear that to investigate the test alone is not enough to support the score interpretation, the relationships between the constructs of interest in the context of the test (underlying notions of skill in algebra and calculus, particularly the role of these skills in the curriculum of the calculus course) and intended placement decisions must also be supported with evidence and rationales. Kane suggests the kinds of evidence which might be offered to support each group of assumptions. The value of the example is in its illustration of the principles that Kane promotes. The disadvantage is that the case is not real, nor did Kane actually conduct the studies. This means that practical evidence for hidden assumptions, inter-dependencies, and long time lines for producing evidence is missing. Such practical limitations are very important for test development boards.

Kane's proposal for the structure of a validity argument is nevertheless stimulating because it offers a clear model for what modern validity theory suggests test developers, validators, and/or users should do. However, it is very clear that the interpretive argument is focused on defending one contextualised interpretation. The case is built when a proposal is made to interpret scores in a certain way. If a different interpretation or use is proposed for the scores from a test, a new argument should be built. Some parts of it may be similar to the earlier investigation, other parts, especially those relating to the new use, will be different. Furthermore, as discussed above, although the logic of Kane's proposal is appealing, he has not published an actual validation case which would follow the model.

The 1999 *Standards for educational and psychological testing* (AERA 1999:9-11) implement Kane's proposals in their recommendations for validation practice. The introduction to the standards on validity states

that validity "logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use". The construct is then described in detail, including its relationships with other possible constructs. Next, the validators should identify the types of evidence that will be important for validation, which happens through a set of propositions that support the proposed interpretation for the purpose identified at the beginning. It may help identify the propositions to consider rival hypotheses that might challenge the proposed score interpretation. Once the propositions have been identified, "validation can proceed by developing empirical evidence, examining relevant literature, and/or conducting logical analyses to evaluate each of these propositions" (AERA 1999:10). Like Kane and Shepard, the *Standards* also point out that "strong evidence in support of one [proposition] in no way diminishes the need for evidence to support others" (AERA 1999:11).

3.5.4 Research techniques employed in validation

All the frameworks presented above offer guidelines on what to investigate in the process of validation inquiry, and how to organise the inquiry. They also mention research techniques for validation, which are discussed extensively in several theoretical presentations of validity. Cronbach (eg. 1990) and Messick (eg. 1989a, 1989b, 1995a), for instance, provide comprehensive discussion of possible techniques. A very similar range of techniques in the context of language testing is presented in several articles in Clapham & Corson (eds.) (1997), which give concrete examples of research with language tests where the techniques have been used.

Because of the history of the validity concept and the current focus in validity on score interpretations, many of the validation techniques are numerical and use test scores as the primary data. Correlation, however, encompasses only part of the analyses and indices involved. Internal correlations to discover relationships among items, and external correlations to investigate the relationships between the scores and other indicators of interest, are the most obvious. Connections to underlying dimensions which might stand for ability constructs are most commonly made through factor analysis. Score-based investigations, however, also include generalizability studies of interpretations across items, populations, and raters. Such studies take into consideration the test, the interpretation, and the population of test takers. Additional test-taker related investigations include the stability of scores over time and across different subgroups of test takers. Insights into the construct measured can be gained by keeping the group of test takers

constant and altering the testing conditions. Alternatively, the testing conditions can be kept constant, but the test taker group manipulated, for instance through providing extra training in the skills measured, or in test-taking skills; the latter in order to see if the scores can be influenced by coaching. Recent advances in quantitative validation techniques are reviewed in several chapters in Clapham & Corson (eds.) 1997, eg. Bachman, Bachman and Eignor, McNamara, and Pollitt).

Proposed techniques which use something else than scores as their primary data can be divided into three main groups: investigations of test content, processing, and test discourse which is assessed to arrive at the scores. The sole technique suggested for the validation of the content of a test is expert judgement. To elicit the judgements, test developers must produce a domain specification and a test specification against which the actual test forms can be judged. The judgements should concern both the relevance and the representativeness of the content of the test.

Techniques recommended for the analysis of processing are more varied, including think-alouds, retrospective interviews, questionnaires, computer modelling, and experimental control of sub-processes. Studies employing such techniques focus on the nature of the construct and the actions through which the assessment is realised, ie. Messick's (1995) substantive aspect of validity. These studies often concentrate on test taker processing, which is understandable, because the score is assigned to the test taker, and it should say something about the test taker's skills. As Banerjee and Luoma (1997) note, however, assessor processing has also begun to be investigated in tests which rely on human assessors.

Moreover, the language samples which the raters rate are also beginning to be analysed to provide an additional perspective on what it is that is being assessed. In language testing, such studies concern assessment of writing (eg. Bardovi-Harlig and Bofman 1989, Ginther and Grant 1997) and speaking (eg. Lazaraton 1992, 1996, O'Loughlin 1995, Ross 1992, Young 1995). The researchers usually approach their data from a conversation analysis or discourse analysis perspective, and often count and describe interesting instances of language use. Apart from characterising test discourse and facilitating construct-related inquiries of how it compares with non-test discourse, these studies offer useful material for testing boards because they allow them to assess the quality of their test and its implementation. They can investigate, for instance, whether the scale descriptors that they use actually correspond to the features of discourse found in the performance of examinees who are awarded each of the scores. They can also study whether the interlocutors act as instructed and

in a comparable manner with each other. The results may lead to minor or more major revisions in the testing procedures.

A research technique which combines qualitative and quantitative analyses for validation is the range of judgmental methods employed in standard setting. If a language test uses reporting scales to assist the interpretation of test scores (e.g. reports the scores of a reading test on a five-band scale), a very important aspect of its validation is the setting of the cut points that divide the distribution of scores into categories. This is usually done by having experts give judgements on items or learners. The experts should be well qualified for their work and the procedures should be well enough specified to enable them to "apply their knowledge and experience to reach meaningful and relevant judgments that accurately reflect their understandings and interpretations" (AERA 1999:54). One such procedure was specified in the context of the DIALANG assessment system by Kaftandjieva, Verhelst and Takala (1999). In it, qualified experts were trained in the use of the Council of Europe descriptive scale (Council of Europe forthcoming) and in a highly specified procedure judged the difficulty of each item that had been pretested. The judgement information was combined with empirical difficulty information from piloting to set cut scores. Such procedures provide empirical evidence for meaningful score conversion from a measurement scale to a conceptual reporting scale and support the validity of score interpretations in terms of the descriptive scale.

3.6 Issues relevant to the present study

The theoretical literature on validity in educational measurement is in agreement that validation is a broad concept and that it involves several actor groups. Test takers, score users, and society that evaluates people partly on the basis of how they fare in tests must bear part of the responsibility for validation activities. However, current literature is less specific on the actual responsibilities of test developers, and it is not particularly common for theorists to look at validation from a test development point of view. Below, I will take up three strands of discussion in validity theory that are particularly relevant for the study of test development. These concern the status of the test in validation inquiry, the role of the construct definition in test development and validation, and the way in which decisions about test design, like research designs for validation, reflect the values of the test developers. I will summarise the existing discussion on these topics and draw the implications in terms of validity theory's recommendations for test-based validation practice.

3.6.1 *The status of the test in test validity*

Cronbach and Messick, and other measurement experts in their wake, strongly emphasize that one does not validate a test but interpretations of test scores. Cronbach, for instance, proclaims:

> Only as a form of shorthand is it legitimate to speak of "the validity of a test"; a test relevant to one decision may have no value for another. So users must ask, "How valid is this test for the decision to be made?" or "How valid are the several interpretations I am making?" Cronbach (1990:150)

Messick (1989a:13) similarly states that "what is to be validated is not the test or observation device as such but the inferences derived from test scores". These formulations shift the focus from the test to score use, and at the same time, they imply several actors. The people responsible for the validation of score-based inferences are both the test developers and the score users.

The mainstream version of current validity theory provides a coherent context for these statements, but it is not easy to implement the statements in one coherent line of validation practice. This has led several measurement experts to call for clear guidelines specifically focusing on the responsibilities of the test developers (eg. Maguire, Hattie and Haig 1994, Shepard 1993, Wiley 1991, Yalow and Popham 1983). The writers make different cases, but they are joined in the concern that current theory and standards on validity do not give sufficient guidance to individual testing boards. Evidence for the quality of a test is only one strand in cases which concern the quality of score-based inferences. For individual testing boards, however, their test is the main concern throughout its development and use. There need not be a conflict, but some clarification of responsibilities is needed.

Yalow and Popham (1983) want the content of a test to be a clear and legitimate focus of validity inquiry. They especially take issue with Messick's (1980:1015) characterisation of content validity as an aspect of test construction and "not validity at all". Messick argued this because content validity is a stable property of the test rather than scores, and does not concern the nature of the skills represented in test responses as validity should. Yalow and Popham see content validity as a necessary precursor to drawing reasonable inferences from the test scores. Messick (1989a:36-42) discusses the contributions and the limitations of content investigations to validity arguments in detail and concludes with an emphatic statement to the effect that content relevance and representativeness do contribute an important perspective to validity investigations. They just cannot be the *only* basis on which a validity argument stands. The implication for test

developers is that content related evidence is important, but it must be complemented by other evidence from the test development process to construct a solid validity case.

Wiley (1991) argues for a return to *test* validity, which he sees to be focused on the social and psychological processes which the test performances are supposed to reflect. The difference between his case and Yalow and Popham's is that Wiley does not consider task or content characterisation only, he focuses on both tasks and test taker processing, which are combined into a particular kind of construct definition for the test. Wiley proposes that test validation should be an "engineering" task which investigates the faithfulness with which the test reflects a detailed model of the intended construct. According to him, this should be kept separate from the "scientific" task of validating the construct model. He presents an approach to modelling constructs as complex combinations of skills and tasks and provides an example of how test validation could be conducted without reference to the scientific validation of the construct. The case is appealing, but it clearly builds on the presupposition that test developers can draw up a detailed model of the intended construct. A particularly attractive feature in Wiley's case is the limitation of the test developers' responsibilities. Shepard (1993:444) and Moss (1995:7), though from a different viewpoint, make a similar case for separating test-related validation from the validation of theoretical constructs.

Maguire, Hattie and Haig (1994) read Messick's (1989a) emphasis on score use to mean that he thinks that the use to which people put a score as an indication of a construct is more important than an understanding of what the construct is, and they disagree. They consider investigations of the nature of educational constructs to be the most important, and they promote qualitative, processing-oriented studies for inquiring into them. They point out that too much interpretation and theorising about constructs in testing is based on *scores.* Such investigations, they argue, conflate the nature of the construct and the properties of the scoring model which is used in the test. Tests can help to build construct theory, but primarily through opportunities for qualitative evidence about test taker processing. Once processing-oriented studies have resulted in a detailed construct, they suggest that educational measurement experts should probably consider whether such constructs can be measured along a scale as current tests do, or whether it might be better to assign test takers to nominal categories which are not necessarily ordered on a single dimension.

Maguire, Hattie and Haig's (1994) proposal holds merit if constructs are to focus primarily on cognitive processing. However, the question can

also be raised whether processing is a sensible primary basis for defining constructs, given the contextual nature of human processing. Furthermore, we know very little about possible variation in processing when one person takes one task on one test occasion versus taking it on another occasion, let alone the differences in processing between individuals – on one test occasion or across different test occasions. In fact, at least some of the evidence available from think-aloud studies, eg. Alderson's (1990:470-478) analysis of the processes that two learners went through when answering a reading test, suggest that variation in processing can be considerable. It is also possible that cognitive processing is significant in some tasks, but that in others, such as in the ways in which readers achieve an understanding of a text, the specific processes employed are neither a sensible nor perhaps a useful way of analysing their skills.

Maguire et al.'s (1994) contribution to the construct validity discussion is nevertheless thought-provoking. The research they promote seems to belong to the theoretical side of test-related and theory-related construct validation, as discussed above. Yet the separation they make between skill-constructs and their quantified indicators is important, as is the question that they raise about whether different score categories justifiably indicate higher or lower levels of "ability". Such a questioning approach would probably be welcomed by Messick and those who continue his work, because it directs attention to the values and practices of current educational tests. Such studies undoubtedly have implications for individual tests, but it might be more justified to see this line of inquiry as a "more scientific" pursuit in the first instance and the domain of an individual testing board's validation activities only after some research basis exists to which they can tie their investigations.

Neither Cronbach nor Messick explicitly discuss the status of the test instrument in their theories of validity. Both theorists, instead, centre validation on the construct which the test is intended to assess. The test tasks, the scoring system, and the score interpretation are referenced to evidence about the construct. But the construct is abstract and related to other constructs and construct theories as well, and construct-related inquiries may end up questioning the construct as well as the test. Test developers may be happy to agree in theory, but in practice they face the question of how to implement a focus on the construct in their validation work while keeping the scope of their task in manageable proportions and continuing to develop and implement their test.

3.6.2 Construct theory and construct definition in validation inquiry

Validity theorists may be unanimous that the construct is the central concern in validation, but it is not entirely clear how this can be implemented in test development and validation practice. Theorists do criticise current implementations. Cronbach (1989:155), for instance, complains that test manuals "rake together miscellaneous correlations" when they should "report incisive checks into rival hypotheses, followed by an integrative argument".

The reason for the current state of practice may be that test developers do not have clear models for how to ground validation inquiry in construct validity. They focus on reliability, item homogeneity, and sometimes prediction of specific criteria, because the procedures for providing these types of evidence are clear. In contrast, advice and examples of how constructs should be defined and especially how these definitions are to be used in validation is largely missing. Furthermore, advice on linking the activities of test development and validation is also largely absent.

I would like to suggest that the first step to make the construct central for validation is to start the inquiry from characterising the construct. It is possible that this point is so simple that theorists assume it automatically, not giving its implementation much emphasis, but unfortunately this makes it easy to overlook this step. As was discussed in Chapter 2, data on the construct definition is available from the test development process provided that test specifications are written and that the steps of test development and the reasons for changes in the specifications and draft tasks are recorded and considered from the construct point of view. Furthermore, I propose that in order to implement validation in the way it is currently presented in theoretical writing, all subsequent stages of validation should be referenced to the construct definition or refined versions of it. If this is not done, the grounding rationale for validation inquiry must be sought elsewhere and if the solution is to ground it on the numerical values of the scores, the test developers may be left with "raking together miscellaneous correlations" which are not connected through a construct-based argument. Moreover, given that the construct definition is also a central guideline for test development, this solution also offers a way to tie the processes of test development and validation closely together.

Some advice for the construct definition *can* be found in current validity theory, especially in Cronbach's writings. He seems to think (1989:151-152) that test developers shy away from attempting to define their

constructs because they are not theoretically solid, and he provides help: "A test interpretation almost never has a consolidated theory as its armature; mostly, we rely on crude theory-sketches. The loose assembly of concepts and implications used in typical test interpretations I shall call 'a construction' rather than a theory." Cronbach (1990:179) explains that very detailed construct definitions are not necessarily required at the beginning of a test development project; the definition can be refined as development proceeds and in this process, score data forms an essential part of the developing case of construct validation. Nevertheless, validation is clearly focused on score interpretation, which involves construct definitions and construct theories. Construct definition begins from a statement of the purpose of the test. The refinement is based on explicit consideration of the rival hypotheses and explanation of how the construct of the test was related to them. Cronbach (1988:5) admits that this strategy requires the skills of a devil's advocate, and that it is "hellishly difficult," but it is vital for the validation effort that this be done.

The construct validity rationale advises that the relevance of the test construct for the proposed use must be evaluated, and at least judgemental if not experimental data should be gathered to support the case that the test be used. The more clearly a test construct is defined verbally, the easier it is to make evaluations of its relevance for a proposed use. The better the evidence for what high and low scores stand for, the easier the basis for developing and interpreting score-based cases. Use arguments always require a verbal and numerical specification of the intended use as well, of course. If the purpose of the test has been defined carefully, and if the proposed use is close to this, cases should be easy to make and evaluate.

3.6.3 Values reflected in test development and validation inquiry

All through his writing, Messick has emphasized that both meaning and values must be taken into account in test validation. The centrality of this is concern is evidenced through many of the titles of his papers on validity: 'The Standard Problem: Meaning and Values in Measurement and Evaluation' (1975), 'Test Validity and the Ethics of Assessment' (1980), 'Evidence and Ethics in the Evaluation of Tests' (1981), 'Meaning and Values in Test Validation: The Science and Ethics of Assessment' (1989). Meaning in Messick's writing is specifically focused on *score* meaning, and he argues that values are an integral part of score meaning because they are automatically engaged when the scores are interpreted and used in society.

Messick feels that value issues must be handled in a validation exercise because their existence and effects cannot be avoided in score use.

This is so "because psychological and educational variables all bear, either directly or indirectly, on human characteristics, processes, and products and hence are inherently, though variably, value-laden. The measurement of such characteristics entails value judgments–at all levels of test construction, analysis, interpretation, and use–and this raises questions of both whose values are the standard and of what should be the consequences of negative valuation" (Messick 1980:1013).

The value concerns that Messick has raised by including consequences of test use in validation are broad and complex social issues. The fact that Messick raises them highlights the social responsibility of test developers and score users, but the difficulty is that the scope of validation becomes very broad and combines test development, score use, and social policy. Nevertheless, the point stands that value implications are unavoidably involved when tests are developed and used. With respect to test development, it is possible to define a limited selection of the broad issues that Messick raises, however. These would focus on the values that can be seen to have influenced the decisions that were made in test development and in the implementation and publication of certain kinds of validation studies and not others. These questions are not as complex as those quoted above – whose values are standard and what the consequences of negative valuations should be – but they provide a basis for asking such questions, and at the same time they are related to the concrete processes of test development and validation. If the basic validation activities comprise the accumulation of evidence and rationales to support the preferred test interpretation and the relevance of score use in particular situations, as current validity theory holds, the analysis of the values that underlie test development decisions and validity rationales would complement this evidence, specifically from a value perspective. This would not remove the need to provide empirical evidence for the quality of the test, quite the contrary. Empirical evidence can support validity arguments and lack of evidence can fail them. By addressing the value implications, the post-modern realisation that there is no value-free standpoint would nevertheless be taken into account.

3.6.4 Test-related validation: when and how

In the introduction to the present thesis, I mentioned the test developers' practical concerns of "when do we do validation" and "how". The answer to the first question, on the basis of the discussion of validity theory above, is "all the time". The answer to the second question is slightly longer, but its

essence is a call to define the test construct and use it as the principal guideline to organise test development and validation.

According to my interpretation of validity theory, it advises test developers to write down a characterisation of their construct early on regardless of how sketchy the initial wordings may be and keep revising it as development proceeds. Any results from studies on the test should always be checked against the construct definition and necessary changes to the construct definition or to the test should be made and recorded.

Another important point is that developers should address the threats of construct underrepresentation and construct-irrelevant variance. In other words, they should ask: "does our test include all the dimensions of ability that we claim it includes and/or we think it should include"; and: "are the scores from our test significantly influenced by other constructs than the one we have defined". When designing validity studies, they should address possible alternative interpretations of the test scores. When assessing the comprehensiveness of their validation efforts, they could use Messick's or Chapelle's six components of validity inquiry or the five sources of evidence from the current *Standards for educational and psychological measurement.* These cover the test content, test taker processing, the way the construct is implemented in the test items, the test's relationship to other tests and constructs, and the likely or claimed consequences of using the test. All of these activities are centrally connected to the construct which the test implements. When they assess the values implemented in their development and validation processes, the test developers should study what they have decided to assess, how, and why, and what kinds of questions they ask in their validation studies and what data they use in the investigation. A review of the answers would prepare them for a social discussion about the political and power dimensions of their test and perhaps help raise questions about the potential social consequences of the use of their test.

4 APPROACHES TO DEFINING CONSTRUCTS FOR LANGUAGE TESTS

In this chapter, I will discuss current approaches to defining constructs for language tests. The purpose is to summarise the advice from existing literature and to illustrate the range of alternatives that test developers can use to define the constructs that their tests assess. To make the discussion concise, I will focus on the range of approaches available and only present one or two prominent examples for each approach.

I will take up two main kinds of approaches: theoretical and empirical. Theoretical approaches help the test developers create a conceptual network that guides score interpretation, the need for which was identified in the previous chapter. Empirical approaches use data from tests in order to investigate the nature of the construct assessed in the concrete case of a single test. In discussing the examples, I will address the following questions:

- What is the nature of the constructs that the different approaches define?
- How are the constructs related to, and reflected in, the test instrument, the testing process, and the test scores?
- How can test developers use the different approaches to construct definition in test development and validation?

I will begin the chapter with a brief summary of reasons why language testers should define the constructs assessed in their test. Next, I will show an example of how this is (not) realised in the practice of test evaluation. In preparation for answering the questions above, I will briefly discuss two models that distinguish elements in the operationalization of constructs, one that describes the influences that an interactionalist construct definition hypothesizes to underlie performance consistency in tests, and one that depicts language testing as an interactive event. Then I will discuss theoretical and empirical approaches to construct definition. I will take up three types of theoretical models of language ability, which are all interactionalist in orientation but which differ in terms of what aspects of ability they focus on: componential models, performance models, and processing models. This will be followed by a treatment of empirical approaches to construct definition, which differ in terms of the materials and methods they entail. I will conclude the chapter with a summary and a discussion of the use of construct definitions in test development and validation.

4.1 Reasons for construct definition

Test developers need to be able to say what the test scores mean. In this sense, construct definition is an accountability concern. What this means in practice, how detailed the descriptions and definitions need to be for different purposes and audiences, and what evidence for the score meaning is necessary or sufficient are questions which have no standard answers; test developers must find their solutions in their specific contexts.

Constructs have to be defined because the words used to explain what the test is testing guide the score users' generalizations from scores to likely examinee abilities in non-test language use situations. Some of the approaches to construct definition in language testing identify a few components of language ability which allow a fairly broad generalization because the components are hypothesized to be central to a wide range of communication situations, while other approaches are quite narrow and the authors warn that generalizations should only be made with caution. However, no test developer or researcher who is actively involved in language testing could recommend the use of no generalization at all; if the test only indicated the participants' performance on the test and said nothing else about their ability, there would be no point in testing. The nature of the construct definition used is likely to guide the nature of the generalizations intended and supported by the test developers. The most common overall constructs currently used in language testing are overall language ability and the skill constructs of reading, writing, listening, and speaking. A typical generalization might concern eg. ability to speak English in teaching contexts.

However, the constructs which are needed in test development are more detailed than the generic interpretive constructs. Test developers need to provide guidelines for item writers about what the items should test and they need to develop assessment scales to regulate how the assessors will assess performances. These more detailed constructs presumably have an influence on what the scores mean, or at least they are used in the examination to create consistency between test forms and assessment procedures. The relationship between this consistency and the construct assessed in the test is an interesting object of study. It is also arguably a key criterion in the assessment of test quality, at least internally to a test development team.

It can be argued that a commitment to construct description is useful for test developers because it encourages self-monitoring and the improvement of professional activities (Alderson 1997). All tests necessarily embody a view of language, language use, and language

learning because they implement language use in some way in their tasks and define a dimension of ability through their scoring systems. Expressing the view in words makes it more conscious and thus more easily available for examination and improvement. All the frameworks of test development discussed in Chapter 2 rely on construct description to bring coherence to the assessment instrument.

Construct description is also central to validation, especially since validation and construct validation are seen to be more or less co-referential. Davies (1994) quotes a 1970s intelligence tester's distinction between the "old" validity question of "does the test measure what it purports to measure" and the "new" validity question of "just what is it that this test does measure" and calls for a third step, a return to stated test constructs and investigations of whether the tests actually test what they say they test. This is the theoretical grounding of a construct validity argument, and it builds on a detailed construct description.

Furthermore, it could be argued that construct definitions are needed by everyone involved in the process of test-taking and assessment: test takers because they need to know what to prepare for in the test, interlocutors and assessors because they need to implement the assessment in comparable terms, and score users because they need to know how to interpret the scores. However, the needs of these groups of people are different and some of them may not feel any need for a construct description at all. A test taker may not want to prepare; a score user may simply want "a score" and somebody else's recommendation for the minimum acceptable level. This is not to say that test developers need not develop construct descriptions; just that because there is little demand for publicizing the working definitions, it may not happen. While it may be true that test development and validation revolve around a described construct, requirements for publishing or analysing this description are not made as a rule of course even on technical fora, as the following example shows.

4.2 Lack of technical demand for construct definition in test evaluation

Jonson and Plake (1998) conducted a study into the relationship between validity theory and actual validity practices. Their design was longitudinal: they focused on the ways in which the validity standards in five versions of the AERA/APA/NCME *Standards* were implemented in the evaluation of one test, the Metropolitan Achievement Test (MAT), in successive editions

of the *Mental Measurement Yearbooks* (MMY) (12 editions altogether, starting with Buros 1938). In the course of 57 years, the MAT had been reviewed in the *Yearbooks* eight times. Jonson and Plake developed a matrix of the range of the validity standards mentioned in the successive editions of the *Standards* and compared this with the standards which the reviewers had applied when reviewing the MAT.

Jonson and Plake (1998) operationalized the different versions of the *Standards* into two long lists of classes of requirement, one concerning content validity and the other concerning construct validity. Under construct validity, one of the categories which they identified and analysed was called the test's construct framework. The first edition of the *Standards* from 1954 had required of test developers an outline of the construct theory. From the next edition (1966) onwards, the *Standards* called for full statement of the theoretical interpretation and distinction from other interpretations. An account of the network of interrelationships [between different constructs] had been introduced in 1974, and appeared in all the standards from then onwards. In other words, the requirements in the *Standards* regarding the test's construct network are quite substantial and they have been included in this professional code of practice for a long time.

Nevertheless, when Jonson and Plake (1998) analysed the successive reviews of the Metropolitan Achievement Test, they found no mention of the construct framework categories, either as a validity criterion mentioned by the reviewers or as a category of evidence presented about the Metropolitan Achievement Test. What is more, the discussion and conclusion in the article does not draw any attention to this result at all. My interpretation is that researchers and practitioners in educational measurement find the test construct a difficult concept to deal with, and their answer often appears to be to leave it alone and deal with something else.

I find this contrast between a stated need for construct definitions in theory and the apparent lack of demand for them followed by the apparent lack of actual working definitions of test-related constructs both intriguing and disconcerting. I think construct definitions should be written and used in test development and validation. In this chapter, I will discuss a range of theoretical and empirical approaches to defining constructs for language tests. I will analyse them as alternative, possibly complementary ways for language test developers to describe what their tests are testing.

The contrast between demand and non-supply of test-related construct definitions is closely related to a paradoxical contrast between

psychometric, quantitative quality indicators for the measurement properties of tests and the verbal, theory-based quality indicators for the conceptual coherence of the assessment system. The psychometric quality of a test is very important because it indicates the degree to which the observations have been and can be expected to be consistent. The consistency can be associated with performance variation, of course, provided that the conditions of variation can be specified on theoretical grounds. The reliability of scores is important when decisions are based on differences between scores: accountability requires that the test developers must be able to say which differences are "real". However, both of these considerations also entail the ability to say what the scores and the differences between them mean. This requires verbal definitions that are ultimately based on theory and empirical evidence about how the definitions are related to the test, the testing process, and the scores. To clarify the range of variables that theoretical construct definition involves, I will discuss the components or variables that an interactionalist definition assumes to underlie performance consistency. I will also present a model of the features of an interactive testing process, which can help researchers analyse dimensions that can vary when the testing and assessment situations are interactive and potentially variable rather than pre-determined.

4.3 Factors underlying performance consistency: the interactionalist view

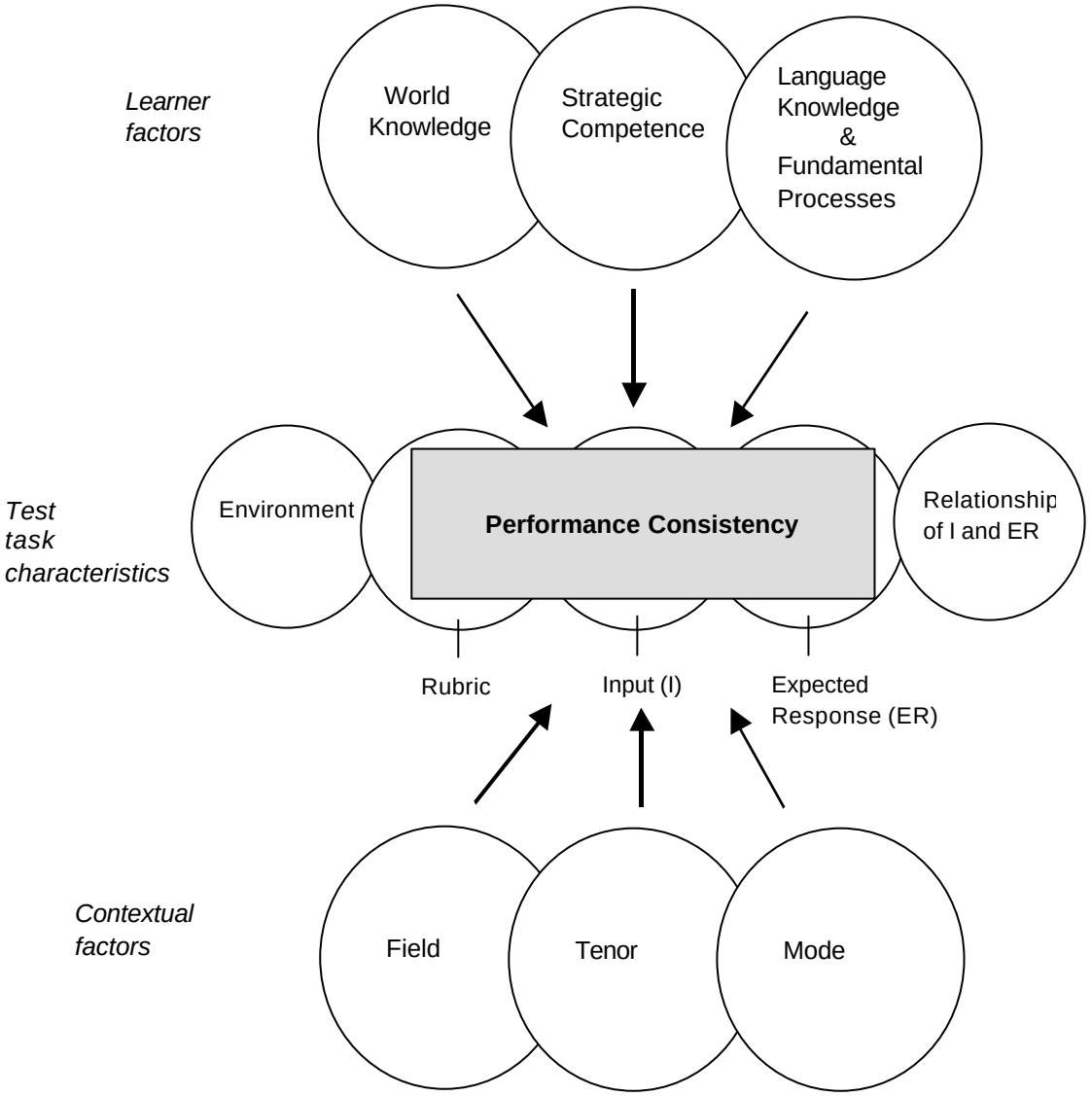
There are potentially a vast range of alternatives for explaining any performance consistencies observed in language tests. Chapelle (1998), following Messick (1981, 1989a), discusses three different theoretical perspectives. Trait theorists "attribute consistencies to characteristics of test takers, and therefore define constructs in terms of the knowledge and fundamental processes of the test taker" (Chapelle 1998:34). Behaviorists, in contrast, attribute consistencies to context. They define constructs "with reference to the environmental conditions under which performance is observed" (Chapelle 1998:34). Interactionalists see performance consistencies as "the result of traits, contextual features, and their interaction" (Chapelle 1998:34). Thus, in order to describe ability, interactionalists would consider it necessary to define the types of knowledge and fundamental processes that the individual has in relation to different contexts and as they interact and vary in response to different contexts. Citing Hymes 1972, Canale and Swain 1980 and Bachman 1990,

Chapelle (1998:43-44) argues that there is strong theoretical support for the interactionalist view in current language testing theory, concerned as it is with individual factors, contextual factors and their interaction. This became evident in Chapter 2 through the range of definitions that theorists considered it necessary to define in test specifications. Chapelle (1998:47) warns that the challenges of this perspective are considerable, however, because it combines two philosophies that locate the explanation of consistencies in different parts of an interactional world: one with the individual across situations, the other with situations or contextual characteristics across individuals. The combination requires the analysis of an individual's abilities in interaction with different contexts. To explain or even detect performance consistency in such a complex network is a complex task. Chapelle (1998:52) illustrates the interactions between its variables with a figure that I will reproduce in Figure 1.

Figure 1 illustrates the range of factors that are required in an interactionalist model of construct-test relationships. Similarly to trait theories, the learner-related factors of language knowledge and fundamental processes are included, but because the learner's interaction in different contexts also has to be modelled, it is necessary to assume that the learner uses strategies to facilitate the interaction and that in addition to language knowledge she also needs world knowledge that varies by situation. Contextual factors are detailed in Chapelle's model with the help of Halliday and Hasan's (1989) theory of context. Their concept of field refers to the locations, topics and actions in the language use situation, tenor includes the participants, their relationship and objectives, and mode includes a definition of the communication mode through channel, texture and genre of language as it is contextualised in the situation. For the analysis of the settings where performance consistencies are sought in tests, Chapelle incorporates Bachman and Palmer's (1996) task characteristics of rubric, input, expected response, and relationship between input and expected response. Chapelle (1998:57) points out that in an interactionalist definition, task characteristics cannot simply be dismissed as error or undesirable construct-irrelevant variance. Instead, researchers and testers must consider some of the contextual variables of test tasks as relevant to the interpretation of performance consistencies.

The approach or model to which a theorist or a test developer adheres is important because it "encompasses beliefs about what can and should be defined, how tests should be designed, and what the priorities for validation should be" (Chapelle 1998:50). The richness of Chapelle's

Figure 1. Factors in an interactive model of language testing (Chapelle 1998:52)



framework for the evaluation and analysis of influences on a testing event may be daunting and all of them cannot be operationalized in an intentional way in a single study or test instrument, but the advantage of the richness is that it enables test developers to focus on different facets in test-construct relationships. At the same time, the complexity presents a warning against simple interpretations of data. In the rest of the thesis, I will use this model as an organising framework to identify areas in which test developers and researchers in language testing have worked.

4.4 An interactive view on language testing

McNamara (1996) and Skehan (1998a, 1998b) similarly argue for the need to see testing as an interactive event. Their point is that the administration and scoring of tests where the participants engage in interaction are activities and that the interactions within these activities have an effect on the scores. Thus, scores cannot be interpreted directly as signs for candidate ability, nor can the external influences on the scores be summarised in terms of superficial features of the test instrument only. This extends the Bachman and Palmer (1996) concept of task characteristics because the interactions of the testing process cannot be controlled in advance in similar ways as the features of tests of receptive skills. The interactions contained in test performance and assessment can also form important influences on the scores, and such influences can be fully justified, perhaps even desirable, in the context of an interactive model of testing. The factors and interactions are complex enough to warrant modelling.

An initial model for language testing as an interactive event was proposed by Kenyon (1992) when he compared selected-response testing and testing which involves performance assessment. His point was that while the score is always a key product of the test, the derivation of the score is more complex in performance assessment than in selected-response testing. This is because, in addition to candidate performance on tasks, the performances have to be rated by raters using scales. From a test development point of view, rating scales are thus an important element of the test, and from a construct point of view, rating scales are an important operationalization of the test construct.

McNamara (1996) extended Kenyon's model by adding the interlocutor as an important variable in task performance. Skehan extended it further still by specifying important features of tasks and task conditions related to candidate processing and by formally adding to the figure the underlying competences of the candidate mediated by ability for use and dual-coding of language. Both authors also discuss rating, Skehan (1998) in passing and McNamara (1996) more extensively, but neither ventured to formally add further variables to the model which would be related to rating or scales. In Figure 2, I have tentatively added two of these variables to herald the discussion of empirical approaches to construct characterisation later in this chapter. Many of these approaches use scales or the rating process as sources of data.

As Skehan states (1998b:84), the model in Figure 2 is helpful in allowing language testers to think about the interactive event of testing and assessment in a more systematic way, so that they can properly consider the

influences other than competence on test scores and so that assessment instruments can be planned and implemented in a systematic and dependable way. From a test development point of view, Skehan's latter point calls for attention to task characteristics and assessment scales in particular.

Figure 2. An interactive view of language testing (slightly adapted from Skehan 1998:172)

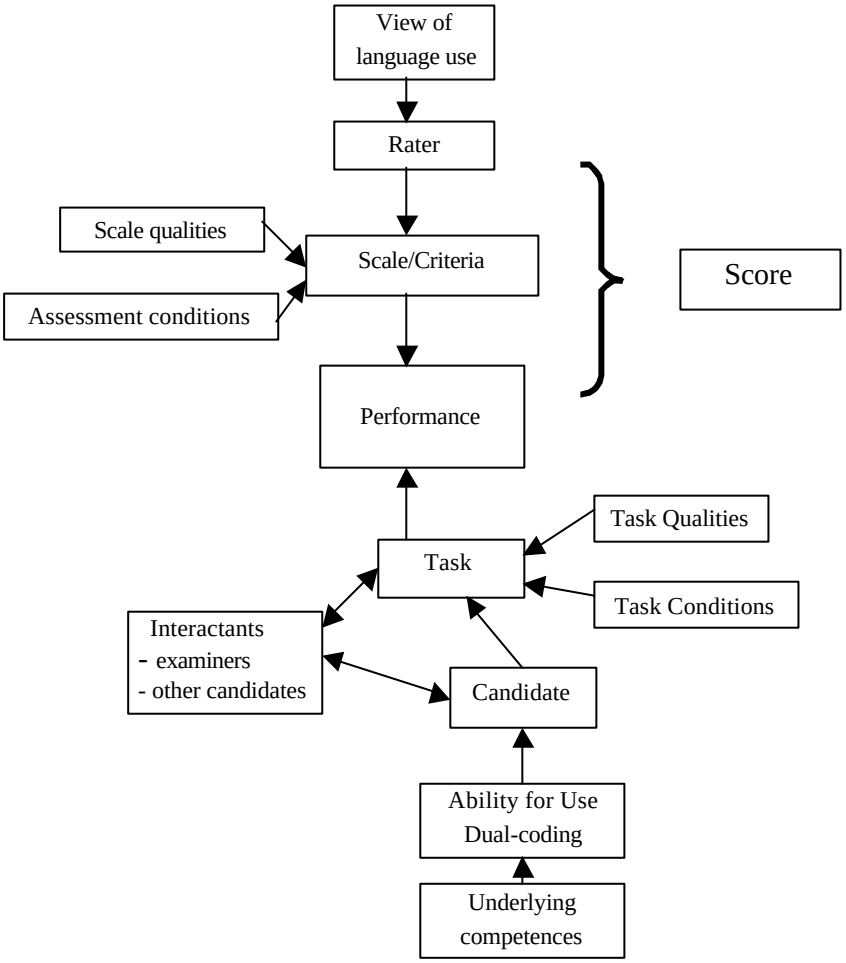


Figure 2 shows that the testing/assessment event implements the test construct in several different ways. Conversely, Figure 2 shows that tests offer several kinds of data from which insights into the test constructs can be developed. The key textual components of the instrument are the tasks and the assessment scales, while the textual products of the testing activity are the test discourse and the scores. The actors in the event are the test taker, the interlocutor when relevant, and the assessor. The key processes are testing/test-taking and assessment, which consist of interactions

between the individuals and the texts or documents. Any of these, or a combination of them, can be used to investigate the construct(s) actually operationalized in a test. Since scores are a key product of the test, the validation-related questions about the test construct and its operationalization are concerned with the different influences which the assessment process has on the scores. The more content information such as diagnostic feedback the score is intended to convey, the more important it is to have evidence about its detailed, construct meaning(s).

4.5 Theoretical models of language ability for testing purposes

Models of language ability and their contribution to language testing have been discussed in detail before (eg. Huhta 1993, Chalhoub-Deville 1997), and I do not wish to repeat what has already been said. Instead, I will discuss the three different perspectives which have been the most prominent in theoretical discussions about language ability in language testing in the past few decades. These are the components of language ability, the processing constructs involved in language use, and the need to model what Hymes (1972) termed "ability for use". To the extent that these theories specify individual skills, contextual factors and interactions between them, they must be considered interactionalist in orientation. It is illustrative to contrast the approaches, however, in order to see the range of differences that can appear within a broad shared understanding of what must be modelled in language testing.

4.5.1 Componential models

Componential models make a distinction between various components in language ability to describe it conceptually. The value of these models for test development is that the salient components they identify can be used as guiding principles in the construction of comprehensive assessment tests and tasks. It is likely that only the parts of a generic model which are relevant for a particular assessment situation will be implemented in any single test, but the comprehensiveness of a componential model can nevertheless support the systematicity of the planning of a test. The model could also be used as a quality criterion for how well a particular test covers whichever areas are relevant for the purpose for which the scores are being used while indicating the areas that are not covered by the test.

The best-known componential model of language ability in language testing is Bachman and Palmer's (Bachman 1990, 1991, Bachman and Palmer 1996) model of Communicative Language Ability (CLA). The CLA model identifies the characteristics of an individual that are engaged

when he or she uses language. These are language knowledge, topical knowledge, and personal characteristics, mediated through, and interacting with, affective factors and strategic competence (Bachman and Palmer 1996:62-63).

Language knowledge in the Bachman and Palmer model is divided into organizational knowledge, which consists of grammatical and textual knowledge, and pragmatic knowledge, which subsumes illocutionary and sociolinguistic knowledge. Each of the components has further subdivisions. Among the other knowledges which are relevant for language use according to the CLA model, topical knowledge comprises the knowledge about the topic that the individual brings to an interactional situation, while personal characteristics are basic features of the person such as sex, age, and native language. Affective factors embody emotional responses to the communication situation, while strategic competence comprises metacognitive organisation and monitoring of the communication situation.

Furthermore, Bachman and Palmer maintain that the nature of language ability must be considered "in an interactive context of language use" (1996:62) rather than solely on the basis of the characteristics of an individual. Thus it is clearly an interactionalist theory. Bachman and Palmer (1996) propose a checklist for the description of tasks that guides a test developer through a close description of the setting and the language characteristics of the task. It includes the setting of language use in terms of physical characteristics, participants, and time of task and a close linguistic description of the characteristics of the test rubrics, task input (or task material), expected examinee response, and relationship between input and expected response (1996:49-50). This implements an analysis of the contextual factors that Bachman and Palmer consider relevant for the modelling of language skills in tests. The authors also promote the use of a similar checklist to analyse which aspects of language ability the test covers (Bachman and Palmer 1996:76-77).

Bachman (1990:81) points out that the CLA model builds on earlier work on communicative competence by Hymes (1972), Munby (1978), Canale and Swain (1980), Canale (1983), and Savignon (1983). Similarly to its predecessors, the CLA model tries to describe the nature of language ability comprehensively and in general terms. The CLA model is thus grounded in theoretical thinking, but the model has also been influenced by empirical results from a multitrait-multimethod study (Bachman and Palmer 1982). A hypothesized model with three traits was investigated, ie.. linguistic competence, pragmatic competence, and sociolinguistic

competence, using oral interview, writing sample, multiple choice, and self-rating as methods. The findings supported a partially divisible model in which sociolinguistic competence was separate from the other two traits. The findings also indicated relatively strong method effects, which is reflected in the current importance of task characteristics in the CLA framework. However, whereas the earlier test method characteristics were considered undesirable influences on scores, the broader range of task characteristics and the inclusion of textual and discourse features indicates a different attitude to contextual factors in the 1996 version of the theory. The only area that is not particularly far developed in the CLA model in relation to Chapelle's (1998:52) model reproduced in Figure 1 is the area of fundamental processes. In their 1996 book, Bachman and Palmer (1996:62) specify that the CLA model is "not … a working model of language processing, but rather … a conceptual basis for organising our thinking about the test development process."

For test development purposes, the complexity and detail of Bachman and Palmer's model yields checklists to characterise the nature of the test and a guideline to develop assessment criteria. To describe the setting of language use, the test developer is encouraged to describe the physical characteristics of the language use situation, eg. location, noise level, and lighting; the participants in their roles, eg. teachers, classmates, friends; and the time of the task, ie.. daytime, evenings, and/or weekends. Similar categorisations exist for the description of other features of the test, including the language of the task, where the test developers describe the grammatical, topical, and functional characteristics of the task material and the expected response. When parallel descriptions are developed for the test tasks and non-test tasks to which the test is supposed to be relevant, the quality of the test can be assessed, and if significant differences are found, the test developers can try to find a better test method. In other words, test developers can use these tools to state what their test tests and to judge its quality.

According to Bachman and Palmer (1996:193), the measurement process, or the process which produces the scores, consists of three steps: defining the construct theoretically, defining the construct operationally, and establishing a method for quantifying responses. The first step is accomplished by describing the construct in detail through the frameworks discussed above. The second step entails writing test blueprints and actual test tasks. The third step comprises the production of a scoring mechanism for the test. For receptive tests, this means defining criteria for correctness and deciding whether binary 0/1 scoring or partial credit scoring is to be

used. For speaking and writing, this involves the creation of the rating scales. For both procedures, furthermore, the test developers need to decide whether to report the test results as they come from the assessment process or whether some score conversion and combination is to be used.

As for the assessment scales, Bachman and Palmer (1996) argue for the use of analytic scales of a specific type, which they call "criterion-referenced ability-based analytic scales" (1996:213). The scale categories are derived from their componential view of language ability and the scale levels are defined in terms of quantity, from 'no evidence of' to 'evidence of complete knowledge of' whatever category is in question (1996:211). The scale for knowledge of syntax (Bachman and Palmer 1996:214), for instance, ranges from "no evidence of knowledge of syntax" through "evidence of moderate knowledge of syntax" with "medium" range and "moderate to good accuracy within range" where, "if [the] test taker attempts structures outside of the controlled range, accuracy may be poor" to "evidence of complete knowledge of syntax" with "no evidence of restrictions in range" and "evidence of complete control except for slips of the tongue".

Bachman and Palmer (1996:211-212) propose this type of assessment scales in contrast to global scales of language ability and argue that their approach has two advantages. Through analytic scales, testers can indicate the test taker's strengths and weaknesses, and such profile scoring reflects what raters actually do when they rate since the features which raters take into account when rating are expressed separately in their scales. Compared with global scales, these are indeed the advantages, but compared with other analytic scales, Bachman and Palmer's approach is quite abstract. The scale is defined without any reference to actual language use situations and the level descriptors include no examples of learner language. The authors state that the introduction to the scale should include definitions of "the specific features of the language sample to be rated with the scale" (p. 213). However, the authors' example of this in the context of the grammar example is as abstract as the level descriptors: "evidence of accurate use of a variety of syntactic structures as demonstrated in the context of the specific tasks (as specified in the task specifications) that have been presented" (Bachman and Palmer 1996:214).

Bachman and Palmer's scale definitions cohere well with their framework for test construction and their model of communicative language ability because the same categories are used. However, from the point of view of a test development board they raise two concerns. The first is conceptual: the view of language learning that these scales implement

seems to build solely on quantitative increase, which test developers may find difficult to accept and certainly difficult to apply in giving detailed feedback. Such issues can only be resolved through empirical investigation, and Pavlou's (1995) study provides one example. He applied the Bachman and Palmer scale for register, which posits that ability develops from no register variation through a good control of one register and an inconsistent control of another, to a consistent control of a range of registers. Pavlou's analysis of learner performances at different levels indicated that the important variable was not the number of registers commanded but the appropriacy and consistency of choice of register for the task setting. If such a modification to the scale were made, it would formalise the effect of contextual factors in performance ratings. It is possible that Bachman and Palmer (1996) did not consider research to be far enough advanced yet to allow such modification. A continued dialogue about different kinds of scales applied to different tests, performances and ability levels might clarify whether scale differences are due to different views of language and ultimately unsolvable through empirical evidence, or whether a consensus could be reached about a practical application of the concept of register to an assessment scale in the context of an interactional theory of language.

The second concern that Bachman and Palmer's scales raise for test developers is a practical worry that scale descriptors which only include brief quantitative phrases, eg. between a "small", "medium", and "large" range of grammatical structures, is too abstract to support agreement between dozens of raters who may be working on their own after initial training. Similarly to the previous worry, this criticism should be substantiated through empirical evidence. An ideal start for such evidence would be for a large-scale examination to implement scales of the Bachman and Palmer type; otherwise, the effort of using two parallel assessment systems might be too demanding in practical terms.

Davies (1996) observes that while language testers often refer to the Bachman and Palmer model, they tend to acknowledge rather than apply it. Chalhoub-Deville (1997:13-14) makes the same observation in slightly more positive terms, contending that the Bachman and Palmer model, like other theoretical models of language ability, can be used to express the extent to which a contextualised assessment instrument covers a context-neutral model of general language ability. In other words, the model is too comprehensive and possibly too abstract to be implemented in its entirety. I will return to Chalhoub-Deville's discussion of theoretical models versus contextualised assessment frameworks later in this chapter.

Extrapolating from the discussion around the Bachman and Palmer model, it seems that componential models of language ability can support detailed construct description and creation of a coherent examination where the theoretical construct definition, the operational definition in task specifications, and the measurement definition in assessment principles all go together. The model does not pose rules for the degree of detail that test developers should use to describe the construct; it offers a support structure that developers can use if they so decide. The model does not require that the construct be made the driving rationale for test development and validation, but it enables test developers to do this if they wish. The basis of generalization that componential models offer for test scores builds on the model's components. In the case of the CLA, these are the language learner's syntactic, textual, functional, and pragmatic knowledges, combined with their personal characteristics and world knowledge and mediated through their affective response and strategic competence. The authors emphasize that it is important to consider these constructs in the context of language use, which they define through task characteristics. These are the important constructs in language testing, according to this model.

4.5.2 Processing models

Processing models of language ability focus on the cognitive processes that people engage in when they use language. The approach is psycholinguistic and closely related to the psychological notions of memory and attentional capacity, which are relevant because language use happens in real time. Language users have limited short-term memory and limited attentional capacity, and while they use language, they are embedded in an interactive situation where both language and other activities are going on at the same time. Processing models try to specify what is going on in the language user's cognitive system, what they pay attention to, and what their language resources are. In the field of second/foreign language ability, processing models tend to be learning-related, which means that theorists particularly focus on learning tasks, especially ones which might be considered to enhance language learning.

The processing approach to language ability for language testing purposes has recently been discussed by Peter Skehan (eg. 1998a, 1998b). The current version of his theory of second language learning builds on previous work in psycholinguistics on attention, noticing, and lexicalised processing (eg. Nattinger and DeCarrico 1992; Pawley and Syder 1983; Robinson 1995; Schmidt 1990, 1993; Van Patten 1990) as well as data

from a series of studies by Skehan and Foster (1997, 1998, Foster and Skehan 1996, 1997), in which learners were engaged in paired interactions. Skehan and Foster used different task types and varied performance conditions in terms of planning and post-task operations, transcribed the learner performances, and analysed the transcripts through a rough operationalization of fluency, accuracy, and linguistic complexity.

Skehan (1998b) begins his presentation of a processing perspective on second language learning from a theory of attentional priorities in learner performance. Building on Van Patten's (1990) distinction of form and meaning as relatively independent features to which language learners need to pay attention, Skehan makes a further distinction within form between focus on accuracy and focus on complexity. The relative independence of the three factors is given some support by Skehan and Foster's empirical results (1998b:71-72). Skehan presents the case that different kinds of tasks and performance conditions call for different balances of attention to the three factors. Because of this dual focus of individual processing when the individual engages with tasks, Skehan's approach can be considered interactionalist. However, he is clearly concentrated on the *fundamental processes* aspect of learner factors and the processing requirements of tasks.

The task characteristics that Skehan (1998b:79) identifies as important for learner performances are familiarity of information, degree of structuring in task, number and complexity of mental operations required, complexity of the knowledge base that the learner needs to draw on to respond to the task, and degree of differentiation in task outcome. The more familiar the information on which the task is based, the more fluent the performance. Clear sequential structuring in the task, such as narrating a story or giving someone a set of instructions, leads to greater fluency and accuracy in performance than a task which lacks such structuring. If the learner needs to make transformations to the task material, such as combining pieces of background information or creating links between instances, it may require more complex language, but it will reduce the amount of attention available to accuracy and fluency. If a learner needs to consider a complex set of perceptions to explain their point of view, it will require them to use more complex language than would be needed for the expression of simple or clearly structured information. And if there is only one possible outcome for the task, the language that learners use will be less complex than in tasks where there are several different outcome options, any of which would be equally "correct". Skehan (1998b:80)

points out that these dimensions are only likely tendencies, not laws, because there are no hard-and-fast laws in rules of language use.

The differences between Skehan's (1998b) and Bachman and Palmer's (1996) task characteristics are striking. While Bachman and Palmer describe task settings physically in terms of location, duration, or lighting, and task language in terms of its syntactic and textual characteristics, Skehan describes the operations which learners engage in to complete the tasks and predicts their effect on learner language. Yet both sets of characteristics can be used by test developers for the same purpose, namely to categorise tasks for the purpose of covering enough variation in language use within a test and creating parallel tasks for different versions; they result in complementary perspectives to the description of examinee ability.

Skehan's contribution to the modelling of language ability for testing purposes is a closely data-based relationship between competence and performance. Bachman and Palmer model this through various kinds of knowledge (or competences) mediated by strategic competence, but they state (1996:62) that their constructs are not directly related to processing. Skehan, instead, focuses on processing. He takes up McNamara's (1996) model of performance testing and extends it further by including processing-based dimensions to task description as discussed above. Furthermore, he specifies that the candidate's underlying language competence, while probably relevant, is mediated by a dual-coding system for language as well as ability for use. With dual-coding, Skehan means speakers' use of memory-based, lexicalised language as a default in online situations such as spoken interaction, where processing demands are quite high, and their use of syntactic processing for precise and clear expression when the task requires it and processing resources allow it. This follows the work of those who apply cognitive theories on language processing, eg. Pawley and Syder (1983), Skehan (1998a), and Widdowson (1989). The point that Skehan makes is that inferences about learners' language competence based on their performance on a test are mediated by so many factors that interactive, processing constructs might be more useful bases of generalisation. Skehan does not advance new theories for ability for use, but restates McNamara's call that such a model is necessary for performance assessment (1998b:84).

The implication of Skehan's processing model for the characterisation of constructs in test development is a call to pay attention to the processing dimension. In practice, this means that processing-based task characteristics and task conditions should be considered when tests are

being developed. It might also mean that fluency, accuracy, and complexity are used as scoring criteria, at least for oral tests. These are tied to a cognitive processing model of human activity, where the main interest is on attention, performance conditions, and performance features, while the underlying competences of individual learners have a less central role (Skehan 1998a:155). If test developers choose to make processing constructs central to their test, these also form the basis of generalisation for their test results; in other words, they might be able to specify quite concretely the types of language use tasks to which the scores should generalise. However, it would be less likely that they would state the results in terms of the participant's language competence.

Constructs like Skehan's may be most useful in educational contexts as a means for providing diagnostic feedback to learners. It may be more difficult to use these constructs in large-scale examination contexts unless the score users are prepared to accept such highly task-related constructs. Whether test developers choose to align their thinking about the test construct along processing dimensions or not, the existence of such an alternative at least encourages them to think about *why* they mention the kinds of constructs they do and what the relationship is between their constructs, task characteristics, assessment criteria, and scores.

4.5.3 Performance models

The need to pay attention to theoretical models of language performance as opposed to models of language knowledge has been raised strongly in recent years by McNamara (1995, 1996). Theories of performance are relevant because performance assessment has become so widespread in language testing since the rise of the communicative view of language in the 1970s. The need for theory development is urgent, McNamara maintains, because the abilities which enable a person to do well on a performance test are not solely language-related. To account for the meaning of the scores in a responsible way, performance test developers should be able to state which abilities apart from language knowledge their test rewards and be able to show empirically that this is the case. To do this, McNamara proposes that test developers need a theory which specifies what the ability to use a language entails.

The terminological and conceptual source that McNamara (1995, 1996) uses to raise the point about performance models is Hymes's (1972) theory of communicative competence with its two theoretical components of *language knowledge* and *ability for use.* Hymes posits that both of these underlie any actual instance of language use. McNamara (1995, 1996)

reviews a number of recent theories of language using Hymes's conceptual distinctions and pays particular attention to the way in which ability for use is portrayed in the existing models of language ability. He discusses Canale and Swain's (1980) and Canale's (1983) model of communicative competence and Bachman and Palmer's CLA model at some length. He signals the interesting but underdeveloped and slightly contradictory notion of strategic competence in Canale and Swain's work as a good start and discusses in detail Bachman and Palmer's development of this and other work.

McNamara (1995, 1996) considers a working version of the CLA model which stems from a time before the publication of Bachman and Palmer's 1996 book and indicates that he misses personal characteristics (1996:74, 86), which are in fact included in the published version of the CLA model. These are important, according to McNamara, because candidate performance in spoken tests in particular is likely to be influenced by factors such as the interactants' sex, age, or race. He cites research which was beginning to appear on this issue. Furthermore, McNamara welcomes the inclusion in the CLA model of world knowledge, strategic competence, and particularly affective factors, which have not been included in previous models of communicative language ability although they obviously influence test performance. He makes the point, however, that affective factors should be considered more extensively than Bachman and Palmer's point that as positive an affective atmosphere in the test as possible should be created (1996:74). The area is difficult, he acknowledges, and his proposal for ways forward is to investigate the interaction between candidates and interlocutors rather than merely concentrating on an individual candidate (1996:75). To help develop the research agenda, he suggests collaboration with related research areas such as communication studies and behavioural science (1996:84).

McNamara likens attention to performance models to the opening of a Pandora's box, thus indicating the complexity and pervasiveness of the questions that performance models raise. However, his point is that if performance tests are used, the need for models of performance cannot be ignored (1996:85). He criticizes the 'Proficiency Movement', as exemplified for instance by the American Council on the Teaching of Foreign Languages Oral Proficiency Interview (ACTFL OPI) (ACTFL 1986) or its Australian counterpart, the Australian Second Language Proficiency Rating (ASLPR) (Ingram 1984) for requiring actual instances of performance, the interview itself, while asking the raters to only assess knowledge of grammar and native-likeness of accent and refusing to deal

with ability for use. The threat, McNamara contends, is that if test developers define 'ability for use' out of existence, they deny the need to investigate construct validity at the same time. Thus, McNamara's case is that construct description is a necessary step at the beginning of a construct validation exercise, and when the test is a performance test, the human abilities which are required when language is used in interaction are relevant to the test construct.

McNamara (1996:85-86) argues for the need to develop the existing beginnings of models of performance. According to him, a weakness in the current models is that they are individually centred and embody a rather static view of communication. He encourages testers to expand their view towards an activity-based view of language in social interaction. This is closely linked to the area of assessment that McNamara discusses in the book, which is performance testing, mostly the assessment of spoken interaction for occupational purposes. McNamara (1996:86) begins the construct-broadening work by presenting an interactive view of the assessment of speaking (see Figure 2), where the rating is the result of numerous interactions: the candidate and the interlocutor interact with each other and with the test task to produce a performance. Then, using a rating scale, a rater interacts with the performance to produce the score. Each of these interactions has to be accounted for to explain the score, and this cannot be done with reference to language knowledge alone. McNamara calls for a research agenda to investigate the influence of these interactions on test scores. Chapelle's (1998) interactionalist perspective on test performance (see Figure 1) is a further step in this development. The area to which Chapelle (1998) applied her model was assessment of vocabulary. The more complex model of the testing event entailed in the assessment of oral interaction would complexify the interactions further, but some clarification through the use of Halliday and Hasan's (1989) concepts of field, tenor and mode might prove fruitful in the future.

The fundamental problem that McNamara addresses with his study of existing models of performance is one of clear construct description (1996:87). The implication for language test developers is that if their test is a performance test, their construct descriptions ought to take performance and person-to-person and person-to-task interactions into account. How this is done and what it implies for the nature of the construct described is as yet unclear. However, when discussing some empirical approaches to construct description later in this chapter, I will present some attempts made by language testers to analyse interactions.

4.6 Test-based approaches to construct characterisation

The theoretical models of language ability discussed above are intended to give a theoretical grounding to an assessment instrument. The ultimate basis for construct description as presented in such approaches is a theory, some aspects of which are operationalized in an assessment instrument. However, it is doubtful if any test could operationalize all aspects of a theory comprehensively, so the question of what it is that a particular test is testing cannot be answered simply with reference to a generic model. Instead of or in addition to starting from theory, some researchers have proposed ways of characterising the test construct on the basis of the features of the assessment instrument.

4.6.1 Construct characterisation based on score analysis

Chalhoub-Deville (1995, 1997) studied the construct of "proficiency". She proposed that an important point in the testing of speaking proficiency in learning contexts is that the scores should reflect generic perceptions of proficiency, not only the teacher's. She considered this important because "N[ative] S[peaker] teachers, who usually evaluate learners' L2 oral proficiency, are acting as surrogates for the nonteaching NSs, it is necessary to validate these teachers' criteria with those of nonteaching NSs" (Chalhoub-Deville 1995:258). She proposed that the "end user's" perceptions of proficiency can be investigated empirically by asking groups of naïve raters to rate some speaking performances and then analysing the scores which they give. Since native speakers are not a unified group but differ from each other in terms of cultural background and experience of learner speech, it might be relevant to sample several subgroups among native speaker judges.

Chalhoub-Deville makes a distinction between theoretical models, such as Bachman's CLA, and operational assessment frameworks which, in tests of speaking at least, are embodied in scales and scores. She maintains (Chalhoub-Deville 1997:11) that "when the purpose for which the model is to be used is clearly delimited, a [scale-based] parsimonious model, which relates to a theoretical model, but only includes the contextually salient components, is more appropriate". Therefore, when test developers have properly defined the purpose of their test, the language ability that they want to assess, the proficiency level that the test is intended for, and the tasks they are going to use, Chalhoub-Deville (1997:11) suggests that they should empirically derive a specific, contextually appropriate assessment framework for the instrument rather than assume that a generic framework is appropriate. A more specific model would enable them to say more

clearly exactly which variables best explain the test scores and the differences between them. This consideration is important if the scores from a test do indeed vary by tasks and rater groups. Chalhoub-Deville conducted a study to investigate this.

Chalhoub-Deville (1995, 1997) studied "the components employed by native speakers of Arabic when assessing the proficiency of intermediate-level students of Arabic on three oral tasks: an interview, a narration, and a read-aloud" (1997:12). From the performances of six learners, she extracted two-minute samples on each task and played them to three assessor groups, 15 teachers of Arabic in the United States, 31 nonteachers resident in the United States, and 36 nonteachers living in Lebanon. The assessors used a rating instrument which included the overall impression and "specific scales, encompassing intelligibility, linguistic, and personality variables. Some of these scales, such as grammar and confidence, were common across all three tasks and some were task-specific, such as temporal shift in the narration and melodizing the script in the read-aloud" (Chalhoub-Deville 1995:261). The researcher had arrived at the list of the criteria through an analysis of previous research and a pilot run with a working version of the scales. The judges gave their ratings on a 9-point scale from 1= lowest performance level to 9=educated native speaker.

Chalhoub-Deville (1995, 1997) used multidimensional scaling and linear regression to analyse the ratings and interpreted the derived dimensions in terms of the names of the criteria which seemed to belong to the same factor, backed up by an analysis of the features of performance which seemed to have caused the ratings. The results indicated that the proficiency ratings given on each of the tasks were influenced by two main factors, but that the nature and the weightings of the factors varied across tasks and rater groups. Teachers in the United States emphasized appropriate vocabulary usage in an interview performance, creativity in presenting information in narration, and pronunciation with a minor emphasis on confidence when they rated read-aloud. Nonteachers resident in the United States emphasized grammar-pronunciation and appropriate vocabulary use in the interview, creativity in presenting information when they rated narration, and confidence on the read-aloud task. Nonteachers resident in Lebanon emphasized grammar-pronunciation in the interview, grammar-pronunciation with a minor emphasis on creativity in presenting information on narration, and confidence in read-aloud. Chalhoub-Deville did not express the size of the differences in terms of learner scores, ie.. she did not report whether learners scored differently on different tasks and

whether it was possible to combine the information from the different tasks to provide an overall score. Instead, she concluded that oral ratings are context-specific and influenced by both tasks and rater groups. She stated that the implication for researchers investigating oral proficiency was to take care to employ a range of tasks and rater groups, as this would lead to a better understanding of the proficiency construct (1995:275). The implication of her results for test developers, Chalhoub-Deville suggested (1997:17), was that empirical investigation of end-user constructs is prudent especially if the scores are used for making high-stakes decisions. She stated that an advantage of her approach is that it can be employed during the test construction stage, before scores are actually used to make decisions (Chalhoub-Deville 1997:17).

In terms of theoretical approaches, Chalhoub-Deville's construct is interactionalist in that it connects the proficiency of the individual with varied task demands and rater perceptions. The researcher addresses performance consistency through the argument that proficiency ratings are context-specific to both tasks and rater groups but she does not provide numerical data on the size of the differences in terms of contextualised proficiency. The analysis is focused on *ratings* rather than language, so that in the context of Chapelle's model (see Figure 1) she can be considered to analyse some features in the middle bar, namely those of performance (in)consistency and the ways in which they reflect learner factors and contextual features in the assessment of speaking.

The value of Chalhoub-Deville's approach is its considered attention to score user perceptions and the empirical grounding of the constructs derived. However, the study was clearly research-oriented and not aimed at developing a test. While the data were assessments (as made by naïve raters without training), the study did not yield assessment scales complete with level descriptors. Chalhoub-Deville did not specify what type of scale the components should inform, though she made the point that scales should be task- and context-specific. Because she was not actually building a test, she did not need to decide *which* end user group or combination of groups was the most relevant for the assessment context and how operational raters could be made to assess the features which were salient to them. The approach provides interesting, empirically grounded information about audience perceptions of task-related proficiency, but for scale construction and score explanation, test developers need to combine this method with others. Moreover, the author says nothing about the relationship between the different task-specific ratings for individual examinees. Nevertheless, the study serves as a reminder that assessment constructs may indeed be

task-specific, and if detailed feedback is needed in an assessment context, this approach to defining task-specific scales might be able to provide such detailed assessment information.

4.6.2 Construct characterisation based on examinee performances

Fulcher (1996b) also makes the case that rating scales in oral tests are operationalized definitions of the constructs assessed. His study concentrates on verbally defined assessment scales as these are used in examinations. He suggests that if scale descriptors are detailed and clearly relatable to actual language test performances, a validation study of the scale provides evidence for score interpretation which is related to the construct presumed to be assessed (Fulcher 1996b:225). He reports on a study in which such a concrete, detailed scale for perceived fluency was developed and validated.

Fulcher (1996b) developed a data-based scale of perceived fluency on the basis of coded transcripts of recorded oral interviews. This was an ELTS oral interview, and the operational ratings provided a criterion against which Fulcher (1996b:212) could judge the ratings from the experimental scale that he developed. To construct the scale, Fulcher initially distinguished six categories of fluency-related features of learner speech. These were coherent with existing research literature on fluency which discusses "surface aspects of performance which interrupt fluency" (Fulcher 1996b:215), covering pausing, hesitation, and repetition/reformulation. However, Fulcher did not deal with surface features of performance descriptively, but coded the instances of the surface features in the transcripts for assumed rater interpretations of the surface phenomena. Fulcher used his own intuition as a rater to derive explanatory categories for the (dis)fluency phenomena and arrived at the following eight explanatory categories: end-of-turn pauses; content planning hesitation; grammatical planning hesitation; addition of examples, counterexamples or reasons to support a point of view; expressing lexical uncertainty (searching for words or expressions); grammatical and/or lexical repair; expressing propositional uncertainty; and misunderstanding or breakdown in communication (Fulcher 1996b:216-217).

Fulcher notes that some of the explanatory categories do not reflect a linear relationship between phenomena, interpretation, and ability. End-of-turn pausing, for instance, is fairly frequent in the performances of both low-ability and high-ability examinees, but not in performances at the intermediate proficiency ranges. However, the pausing occurs in different contexts, and raters interpret it differently in the two cases. Low-ability

examinees pause to ask the interlocutor to take over before the proposition they are expressing is complete because they do not know how to continue, while high-ability examinees pause after completing a proposition to indicate that their turn is complete and the examiner can take over. (Fulcher 1996b:220-221.) The implication for the development of level descriptors for assessment scales is that unidimensional increase of fluency phenomena and decrease of disfluency phenomena is a simplification which does not tally well with learner performances. Closer and more realistic description of pausing in rating scales, for instance, would take the nature and motivation for the learner's pausing into account.

Having taken the multidimensionality of surface phenomena such as pausing into account when the transcripts were coded for the interpretive categories, Fulcher used discriminant analysis to investigate how well tallies of occurrences accounted for the operational ELTS ratings of the population. Only one person of 21 would have been given a different rating if the experimental scale had been used instead of the operational one.

Fulcher concluded that his results to support the usefulness of the explanatory categories and proceeded to use the categories and the transcribed interviews to construct a data-based fluency rating scale with categories 1-5 described and additional undefined categories of 0 (below 1) and 6 (above 5) added. In addition to the surface features and explanatory categories discussed above, he added descriptors of backchanneling to the final scale, because his review and re-review of the tapes in the course of the study indicated that backchanneling increased with higher-ability students, and he hypothesized that frequency of backchanneling would influence ratings (1996b:224).

Fulcher derived a fluency rating scale from the data and investigated its validity and functionality by asking five raters to use it in the rating of three oral tasks (two one-to-one interviews and a group discussion, as described in Fulcher 1996a). The students rated were different from the group that provided the performance data for the first part of the study, but belonged to the same population. Fulcher (1996b:214) used a G-study to calculate rater reliability and assessed the validity of the scale by investigating group differences and conducting a Rasch partial credit analysis on the scores awarded. The reliabilities and inter-rater and inter-task generalizability coefficients were very high, .9 or above (Fulcher 1996b:226), and this led Fulcher to conclude that the scale was able to discriminate between three teacher-assigned levels of general ability. The Rasch partial credit analysis indicated that the cut points for different skill levels were fairly comparable across the three tasks (p. 227). The

researcher concluded that the scale was relatively stable across task types. Fulcher (1996b:228) reports that an examination of the scale in the context of a different examinee population is under way, which indicates that he considers it an open issue whether the concrete descriptions of fluency phenomena are generalizable across different groups of learners.

Similarly to Chalhoub-Deville (1995, 1997), Fulcher (1996b) focused on the relationship between test task characteristics and performance consistency. In terms of Chapelle's (1998) figure (see Figure 1), then, he also worked with concepts in the middle, but unlike Chalhoub-Deville, his intention was to support the establishment of performance consistency in tests of speaking. He used analysis of learner performance as material and ascribed the assessments to the learners' fluency, which combined the features of test discourse and learner factors. His conclusions concerned the notion of fluency in context, which he sought to describe empirically. His contribution to the construct description issue for language testers is a data-based way to develop rating scales, and he argued that through these means testers could provide construct validity data for the examination at the same time. This is done by describing the construct actually assessed in the examination in a rating scale with detailed level descriptors. The contrast is to existing rating scales, which may not be based on any direct observation of learner performances or systematic collation of rater perceptions but armchair theorising which is not supported by critical conceptual analysis or by investigation against empirical data (Fulcher 1996b:211-212). Fulcher (1996b:217, 221) notes that the explanatory categories he used for rater interpretations of the features of examinee speech are inferences which require validation, but the statistical evidence of the usefulness of the categories for the prediction of the overall proficiency ratings lends some support to the plausibility of the explanations. Furthermore, this approach offers the possibility of creating links between language tests and applied linguistic theory by using theory to suggest descriptive and explanatory categories to be used in rating scales and by using the assumptions of links in rating scales to inform studies of language ability or language learning to see if the links are plausible.

The rating scales that Fulcher developed are long and complex compared with the scales that assessment developers are used to seeing. Each level descriptor is more than 200 words long. If test developers choose to use this method to construct their scales, they would have to make sure that their assessors are willing to work with them. This detail in the scales might provide a useful means for a group of raters to agree on

ratings, but whether this is so should be investigated. The strength of the scales is their direct basis on learner data. A weakness might be that scales from different systems may not be compatible, new analyses would always be needed if a new test were developed. Considering the number of stages needed, this approach to scale development is time-consuming, but if the information from the scale can be used to provide learner feedback and it proves that learners find this useful, an important gain might be made. Further study is needed to verify the case.

Similarly to Fulcher, Turner and Upshur (1996) also used examinee performances to construct assessment scales. The researchers worked together with 12 elementary school teachers and built assessment scales for their ESL speaking test tasks with the help of the teachers' perceptions of salient differences in learner performances. The project took the view that operationalized constructs are task-scale units, and since the project used two speaking tasks, they also developed two assessment scales. Upshur and Turner (1999) discuss the implications of their project to language testers' understanding of the processes of test taking and scoring.

Upshur and Turner (1999:101-102) describe their scale-making activities as analysis of test discourse. Both the process that they used for deriving the scales and the nature of the resulting scales were different from standard test development procedures and also different from techniques used in discourse analysis. The scale-making procedure began with the participants agreeing in broad terms on the ability or construct they wanted to measure. The process itself consisted of iterative rounds of three steps. First, each member of a scale construction group individually divides a group of performances into two piles, top half and bottom half. Second, as a whole group, they discuss their divisions and reconcile differences. And third, they find some characteristic which distinguishes the two groups of performances from one another and state it in the form of a yes-no question. The same procedure was applied to successive sub-samples of the original sample so that six levels of performance were identified. The resulting scale took the form of five hierarchical yes-no questions which characterised salient differences in the sample of performances used in scale-making. The two scales, one for each task in the project, were then applied to the performances of 255 students.

Upshur and Turner (1999) used many-facet Rasch measurement with the program FACETS (Linacre 1994) to analyse the two task-scale units on a common latent measurement scale. The analysis was performed on 805 ratings given by 12 raters to 297 speech performances produced by 255 children. It showed that the tasks were not of equal difficulty, that there

were differences of severity between the raters, and that the score boundaries were also different, so that for instance it was easier to earn a 6 on one of the two tasks than the other (p. 95).

Upshur and Turner (1999) also discussed the scales in terms of the features of language they focused on. Both scales employed fluency to distinguish between the highest level of achievement and the next highest level. Both scales also based the distinction between the lowest and next lowest levels on use of the mother tongue. The intermediate levels, however, proved to be distinguished by different features in the two scales, which according to Upshur and Turner's analysis were related to task requirements and possibly the rating processes. On a story retell, where the raters knew the content that the students were trying to express, the levels were distinguished on the basis of the content of the retell performances. On an audio letter to an exchange student, where raters were not able to make such content assessments, they focused on the phonology and grammar of the students' speech (Upshur and Turner 1999:103-104). The authors suggested that rating scales should be task-specific rather than generic, since effective rating scales reflect task demands and discourse differences. They also speculated that such task-specific application may happen even if raters are ostensibly applying a single standard scale to rate performances on different tasks (p. 105).

Upshur and Turner (1999:103) noted that the discourse analysis of performances produced by their scales was not exhaustive. It only identified features of the performances which were the most salient for the main purpose of the exercise, which was to enable raters to distinguish between ability levels. The features identified were also dependent on the nature of the performances used when the scales were created. There probably were other features which also distinguished between levels of achievement but which were not equally salient to the group of raters, and other performances might have included other salient features. The authors also pointed out (1999:105, 107) that the resulting task-specific assessments of achievement pose a problem of how to generalize from task-based assessment scores to any more generic ability estimates. However, the advantage is that task-specific assessments allow the assessors to give expression to the process of assessment, reflected in their project in the way that task-specific assessment strategies featured in the scales.

In terms of the distinctions of theoretical approach into trait theorists, behaviorists and interactionalists that Chapelle (1998) used, Upshur and Turner's conclusions certainly show that they cannot be counted as trait

theorists. The argument for strong task dependency may be interactionalist, but their reluctance to draw conclusions about individuals puts them on a socio-constructivist dimension of it, which is not separated as a clear category in Chapelle's model. The researchers emphasize that the assessment process influences the scores given and that assessments are task-specific at least to a degree, whether assessment scales recognise this or not. Their method of employing rater perceptions to construct a scale provides yet another strategy for test developers to arrive at concrete formulations for explanations of scores. The method is so empirically grounded, however, that it may not suit all formal assessment contexts, especially if generalizations in terms of individual abilities are needed. If scales are constructed in such a strongly data-driven way, they really are task specific. This makes score interpretation in typical assessment contexts difficult, since the purpose of educational assessment is surely not only to categorise learners into six groups on the basis of a one-off task. There may be a useful purpose for such task-based assessments but, being new, it calls for detailed definition.

If individually based score interpretations are going to be made on the basis of this type of assessments, the meaning of the scores must be investigated, for instance by analysing transcripts of learner performances and examining whether they reflect the features of performance named in the scale-defining questions. Another question which might be asked is whether the scales were task-specific because they were developed to be so. This could be studied through employing a more generic six-level rating scale on the same performances and analysing whether there are differences between the ratings. It would be difficult to prove which scale was "more right", but the data might show whether ratings are scale-specific or task-specific.

A related call to investigate test takers' and assessors' models in action is also made by Alderson (1997). He contrasts explicit and formal models of language, as embodied in theories of language ability, with implicit models that teachers, testers, and learners enact when they engage in language learning, teaching, and assessment. Extending this logic towards test development, this call would also encompass taking account of the models of language which test developers work by when they develop a test. One possible way of gathering data to study the usefulness of such an approach would be to keep account of decisions made in test development. The advantage of such analyses, as Alderson (1997) argues, is that they throw light on the perceptions actually involved in a concrete assessment instrument and its socially used products, the scores and their

interpretation. Once such data exists, judgements could be made about whether perceptions vary and whether it matters. If data is not gathered, the assumption is automatically that it does not.

4.6.3 Construct characterisation based on task analysis

Upshur and Turner's, Fulcher's and Chalhoub-Deville's proposals for the use of data to derive constructs work in contexts where learners produce an extended response which can be analysed and/or rated in detail. However, tests with structured tasks and limited responses, such as those of reading or listening, do not offer such data. Instead, what can be analysed in these settings is the task material. Freedle and Kostin (1993a) report on a study in which they analysed the construct assessed in the TOEFL reading test by analysing the format of the texts and items and seeing how these influenced item difficulty. McNamara (1996), from a slightly different perspective, describes a number of studies which used task content for the mapping of the abilities (ie.. construct) assessed in a test. Both of these methods are *post hoc* explanations in that the dependent variable is item difficulty. However, the advantage is that the task analysis yields a content description of the ability assessed.

Freedle and Kostin (1993a) conducted an analysis of TOEFL reading tests to explain item difficulties and through them the nature of the information yielded by the scores. They investigated three categories of TOEFL reading items: main idea, inference, and supporting idea items (1993a:145). There were 213 of them altogether and they were related to a total of one hundred reading passages.

Freedle and Kostin (1993a) reviewed existing research to assemble a set of variables which had been found to be related to item difficulty in reading comprehension and investigated how well these variables helped predict the difficulty of sampled TOEFL reading comprehension items. The variables characterised the reading passages, reading items, and passage-item overlap. They included features such as number of words, number of negations, location of focal information within passage, subject matter of text, type of rhetorical organization, and frequency of fronted text structures such as cleft sentences. A total of 65 variables were included in the analyses as well as 6 to 11 text-by-item interactions depending on item category (Freedle and Kostin 1993a:146-154). The categorisation into text-related, item-related, and text/item overlap-related features was made because criticisms had been presented that multiple choice tests of reading assess reasoning skills related to understanding the item stem and options rather than assessing passage comprehension. Since the TOEFL test is

based on multiple choice questions, this would be a considerable criticism of the test.

Using stepwise linear regression, Freedle and Kostin (1993a:161-162) found that eight of the variables could be considered significant predictors of the difficulty of the sample of TOEFL reading items that they investigated. Six of these variables were related to the reading passages, two to passage-item overlap, and none to the textual characteristics of the items alone. The portion of variance explained by the variables was 33 percent. The researchers also conducted separate analyses for a non-nested sample of items, that is, a sample of items where only one item per reading passage was included. In this analysis, eleven variables accounted for 58 percent of the variance of the scores. Ten of the 11 variables were related to the reading passages or to passage-item overlap, and one (number of negations in correct answer) to item-related variables.

Freedle and Kostin (1993a:166) concluded that their results supported the construct validity of the reading test, because they were able to show that candidate scores were significantly related to features indicating text comprehension rather than to technical or linguistic features in the items. In a related ETS publication the researchers report that they also found a tendency in the data that the proportion of variance explained was higher for the two lower-scoring ability groups than for the higher-scoring candidates (Freedle and Kostin 1993b:24-25). They suggested (Freedle and Kostin 1993b:27) that think-aloud protocols might be used to clarify the strategies employed by high-scoring candidates, so that item difficulty could be better predicted for them as well. They did not speculate what such variables might be. Their use of the word "strategies" and the method of think-alouds may indicate a suspicion that reader-related variables which concern the *operations* that the items make readers perform could explain further portions of item difficulty for high-scoring candidates. Such reader-related processing variables were not investigated in Freedle and Kostin's study. If these kinds of variables are considered important for a comprehensive picture of the construct assessed, as they might in an interactionalist definition of test-based reading, the degree of variation not explained might be a good result (cf. Buck and Tatsuoka's results discussed below). However, the operationalization of such variables would require careful work before assessments could be made of whether it explains score variation in a systematic way.

Conceptual issues are only one possible explanation for why Freedle and Kostin (1993a) were able to explain item difficulty better for the non-nested sample of items and for lower ability levels. Boldt and Freedle

(1995) re-analysed Freedle and Kostin's (1993) data with the help of a neural net, originally in the hope that this more flexible prediction system would improve the degree of prediction achieved (Boldt and Freedle 1995:1). They found that the degree of prediction did improve in some samples of items, but the variables that the neural net used for the successful predictions were different from those in the linear regression. Only two of the variables, "number of words in key text sentence containing relevant supporting idea information" and "number of lexically related words in key text sentence containing relevant supporting idea information" were the same (Boldt and Freedle 1995:15). The researchers also found that the highest improvement in prediction of difficulty concerned the nonnested sample of items. However, they studied an alternative explanation for this. They drew another sample of 98 items from the 213 and used the same 11 variables that they had used with the nonnested sample to study the degree of prediction. They found that percentage of variation in difficulty explained for this set was almost as high as for the nonnested sample even if the predictor variables were not formed specifically for the new sample. This argued for the alternative explanation that the difference in degree of explanation in the original Freedle and Kostin (1993) study was not due to the independence of the items but to the smaller sample of item difficulties that had to be explained, introducing randomness in the nature of the sample that allowed capitalization on chance (Boldt and Freedle 1995:14). Similarly, although the Boldt and Freedle study repeated the Freedle and Kostin (1993) finding that item difficulty was best explained for the lowest ability groups, it was possible that this was because the number of predictors that were used for that group was the highest (Boldt and Freedle 1995:14). The authors continued that this alternative explanation was supported by the fact that the accuracy of prediction for all the ability levels in their study reflected the number of predictors. The effects of skill level and sample size were confounded and if a design were developed to investigate the cause, Boldt and Freedle proposed that fewer predictors and more items should be used so that the issue could be resolved (Boldt and Freedle 1995:15).

This example illustrates that findings in empirical studies may be explained by the methods used. Boldt and Freedle (1995) seem keen to find a small number of generalizable constructs, since they say that they would like to find few predictors that work across a large sample of items. Another way of pursuing research on this would be to make parallel small samples and investigate how the variables that explain difficulty vary and possibly discover contextual or content-based explanations for *why* they

vary. For both types of research, Boldt and Freedle's (1995:15) observation that ideally such studies would be informed by theory holds true. The nature of the theory would inform the content of the variables studied, or vice versa if theory construction were sought from identifying item properties and using them in prediction.

McNamara (1996:199-213) describes three projects which used what he terms skill-ability maps to characterise the skills assessed in reading and listening tasks. The approach begins from the output of a Rasch item analysis. This locates examinees and items on the latent measurement scale. The researcher then attempts to identify the skills assessed by items at a specific region of the latent ability scale. The logic is that "if the knowledge or skills involved in the items found at a given level of achievement can be reliably identified, then we have a basis for characterising descriptively that level of achievement. If successive achievement levels can be defined in this way, we have succeeded in describing a continuum of achievement in terms of which individual performances can be characterized" (McNamara 1996:200). A researcher who uses this approach thus hopes to be able to say, for instance, that items testing "ability to understand and recount narrative sequence" cluster at one region of item difficulty while items testing "ability to understand metaphorical meaning" would be found in another.

McNamara (1996:201-202) discusses a first language reading test (Mossenson et al. 1987, in McNamara 1996) in which the reading ability scale was developed through the method described above. The scale proceeds in thirteen steps from the identification of the topic of the story through the connecting of ideas separated in the text to inference of emotion from scattered clues. McNamara (1996:205) reports that the validity of the scale has been called to question, both on the grounds that the status of sub-skills in reading is questionable and especially that the methodology used to characterise the content of the items is not reported in the test manual. Further exploration of this method with carefully reported procedures may nevertheless produce interesting results for construct characterisation. The nature of the properties that are ascribed to the items in the reading test is strongly related to the theoretical views of the analyst about what would explain correct or incorrect responses to the reading item analysed.

McNamara (1996: 203-204) also discusses an individual learner map, where a similar mapping methodology was used in a university test of English as a second language, but this time to detail the answer pattern of an individual learner. The basic grid of the map is defined by the latent

ability/difficulty scale on the one hand and the examinee's answer pattern on the other. The logic follows item response theory, which expects that if a set of items is suitable for the examinee, he or she would tend to get items *below* his/her ability level correct, items *at* his/her ability level either correct or incorrect, and items *above* his/her ability level mostly incorrect. Accordingly, the individual learner map includes four regions: easy items which the learner answered correctly, difficult items which the learner answered incorrectly, difficult items which the learner somewhat unexpectedly answered correctly, and easy items which the learner unexpectedly answered incorrectly. McNamara suggests that the last category in particular is useful in educational contexts, because it may indicate where remedial teaching is needed.

McNamara suggests that information from such learner maps might be used in two ways. It could be reported to learners as it is and learners could draw their own conclusions of ability based on their examination of the items. It could also be combined with content analysis of the items to express learner abilities in terms of more general underlying abilities. As with McNamara's earlier reading example, the nature of such abilities would require validation. However, if ability constructs are thought to underlie examinee performance on tests, skill-ability mapping might offer a way to identify and describe them.

As a third example, McNamara (1996:206-210) reports on another Australian test of reading and listening in which skill-ability mapping is used to create ability level descriptors used in certificates. He describes an independent validation study by McQueen (1992, in McNamara 1996) in which the researcher derived from existing research a set of characteristics which could be considered to affect the difficulty of items in reading Chinese. He used the criteria to analyse a test which had already been administered and the scores reported. There was considerable coherence between the factors that McQueen derived and the ones used by the examination. McNamara (1996:210) reports that McQueen's results largely supported the validity of ability mapping in general and at least in the context of the examination. In addition, McNamara proposes that the ability mapping approach could be followed by performance analysis, both of which should operationalize constructs mentioned in the test specifications. This would provide more powerful evidence for the validity of the maps. Such an approach would also strengthen the role of construct definition as a rationale for the development and validation of tests. The approach would allow a wide range of theoretical approaches to construct definition in terms of Chapelle's (1998) model, and since the empirical

logic of the mapping system is based on the IRT ability dimension underlying the scoring system, the connections between the construct tested and the scoring structure would be a natural part of the investigation.

4.6.4 Construct characterisation based on task and ability analysis

Buck and Tatsuoka (1998) describe yet another empirical approach to investigate the ability constructs which can explain test performance. They applied statistical pattern recognition techniques to the analysis of cognitive attributes (knowledge, skills, abilities) which underlie a test of second language listening with open-ended responses. The technique was exploratory and clearly so demanding both technically and in terms of research expertise that operational test development boards with normal funding restrictions would not be able to engage in it as a standard approach to explain what their test scores mean. Nevertheless, since the approach connects test characteristics to person abilities and thus provides another example of what kinds of construct description have been proposed for test data, I will briefly summarise the approach below.

Buck and Tatsuoka's (1998) approach applies rule space methodology, a statistical technique to identify patterns, to language assessment. They explain that in their case, the patterns which they identified were test takers' "knowledge states", which are something like ability maps that show the abilities which each participant has and which he or she does not have. In this approach, researchers begin by specifying the requirements posed by the set of test items investigated in great detail, as in the example of Freedle and Kostin (1993) discussed above. Then they make inferences about the kinds of knowledge and ability which people need to possess to answer the items correctly. Next, they draw a map of each item in terms of which abilities it requires and interpret people's response patterns to a set of items in the light of the evidence afforded by their item responses about the knowledge and skills they seem to possess and the knowledge and skills that they do not seem to possess. If the evidence is not clear, the technique offers a way to indicate this. The approach holds promise, but because the variables identified are cognitive processes, Buck and Tatsuoka warn that its variables and patterns are far less stable than those of exact science.

Buck and Tatsuoka (1998) emphasize the exploratory nature of their study and explain in detail the process that they went through in identifying item characteristics and abilities which account for people's ability to respond correctly. On the basis of a review of existing research and an examination of the actual items involved in their study, their initial pool of

possible attributes numbered 71. Through content analysis, correlations, and regression analyses, they reduced the list of characteristics to 17, and after a first try-out further to 15. Their final analysis took account of the attributes individually and all possible patterns of interaction. These result in individual attributes which, when they co-occur in an item, make it more difficult than the combined results of the attributes individually might suggest. Some 14 interactions were found to be significant. Buck and Tatsuoka's results indicated that 96% of the score variance of 96% of the people they diagnosed was successfully accounted for by the 15 variables and the 14 interactions.

The person attributes in Buck and Tatsuoka's analysis included skills such as the ability to scan fast spoken test, automatically and in real time; the ability to identify relevant information without any explicit marker to indicate it; and the ability to process information scattered throughout a text. The interactions included the ability to recognize and use redundant information, when the response requires more than just one word, and the ability to use previous items to help locate information, when it was necessary to make a text-based inference and the response requires more than just one word (Buck and Tatsuoka 1998:141-143). Such processing attributes would be highly useful for diagnostic purposes if this technology could be used in educational assessments, and they also provide an interesting content angle into the constructs assessed in language tests. The analyses were experimental and technically demanding; whether similar skills would turn out to account for abilities in other listening tests and whether the degree of similarity could be predicted in advance depending on task types used remains to be seen.

An interesting feature in the attributes used by Buck and Tatsuoka was that they were highly contextualised on an item level while also describing person abilities rather than mere test or item characteristics. The construct, given the multiple properties assigned to any item and the open possibilities for interactions, clearly follows an interactionalist logic in Chapelle's (1998) terms. As Buck and Tatsuoka note, external validation of the attributes would be called for. If this proves possible, however, the validated attributes and patterns of attributes could be used to build a model of second language test-taking which is related to both the features of the examination and to test taker abilities.

4.7 Construct characterisation in test development and validation

The two kinds of approaches to construct characterisation discussed above have a common aim, to describe what the test is testing and what the test scores mean, but their perspectives are different. The theoretical approaches start from describing the nature of the construct of language ability. This has implications for what test developers ought to do to relate their constructs to the model. The model and the related procedures of test development specify the kind of meanings that are available from the scores. In addition to relating their constructs to the model, test developers thus need to decide *which* model(s) they want to relate their test construct to and what kinds of meanings they consider useful for score users. The data-based approaches start from a specific research question and a set of data from a test. The results clarify the nature of the construct assessed in the specific test, which has implications for what other test-related constructs might be like. The data-based approaches also offer a specific method for the investigation of test-related constructs in other projects. Examination boards can of course combine both approaches in their activities.

In terms of Chapelle's (1998) model, the theoretical approaches start from the extremes of the model and provide rules and guidelines for the operationalization of the construct in a concrete test and the interpretation of the observed performance in terms of the theoretical construct. The empirical approaches start from the characteristics of test tasks, test performances and the testing process and develop interpretations about what the scores mean in construct terms. The empirical approaches summarized above show that some researchers have concentrated on the nature of the performance and assessment processes in particular in order to study the nature of consistency or otherwise in these interactive events. It is difficult to find a precise location for this research approach in Chapelle's figure, test implementation simply seems to require rather a complex investigation structure to establish that there *is* consistency at all and to investigate its bases. The test development aim in this research is to find justifications for the "right" consistencies and the means for developing them in actual test situations.

The theoretical approaches to construct description discussed in the present chapter are summarized in Table 2. It focuses on the nature of the construct described in the model and the implications for test development activities. All the three approaches covered characterise examinee abilities, but from different perspectives. Bachman and Palmer's (1996) componential model characterises them in terms of conceptual categories,

Skehan's (1998) processing model calls attention to the learner's cognitive processing, and McNamara's (1996) search for performance models emphasizes the procedural nature of testing and assessment.

Bachman and Palmer (1996) and Skehan (1998) both focus on tasks, but from different perspectives. Bachman and Palmer advocate descriptive and linguistic characterisation, while Skehan emphasizes processing demands. McNamara's (1996) focus is on the interactive dimension of tasks, which highlights the fact that the performance is not the examinee's alone, but a co-constructed discourse with an interlocutor.

Table 2. Theoretical approaches to construct characterisation in language testing

Theorist	Nature of construct	Implications for test development
Bachman and Palmer 1996	Construct describes conceptual categories of ability and knowledge involved in language use. Language knowledge includes grammatical, textual, functional, and pragmatic knowledge. In a language use situation, these interact with world knowledge and person characteristics and are mediated through strategic competence and affective reaction to the situation.	Characterise the nature of tasks, because the components of participants' language ability are engaged through them. Provide evidence for score generalization through correspondence of task characteristics between test and non-test language use. Express scores in terms of components of language knowledge and in terms of evidence of degree of mastery.
Skehan 1998	Construct characterises cognitive processing, especially division of attention in real time (spoken) interaction. Different tasks pose different cognitive demands. Effects seen in form and nature of learner discourse: fluency, accuracy, and complexity of learner language.	Start test construction from an analysis of tasks. Analyse cognitive demands of tasks in terms of familiarity, task structuring, range of outcome options, etc. Group tasks on this basis. Assess learners by counting incidences of fluency, accuracy, and complexity.
McNamara 1996	[construct underlying a performance test should model ability for use: theory to cover strategic competence, personal characteristics, candidate-in-interaction]	Analyse test discourse as person-to-person, person-with-task interaction. Analyse rating as rater interaction with performance through scale. Investigate influences on scores. Build theory.

All three approaches also use performance data as evidence of ability: Bachman and Palmer through scores and scales which indicate degree of mastery, McNamara presumably similarly through scores though possibly combined with discourse features, while Skehan promotes analysis of learner discourse with counts of specific features of discourse as indicators of ability. The ability, furthermore, is processing-oriented in Skehan's version, in that language knowledge is mediated by the learner's ability to deal with dual coding, ie. lexicalised and syntactic processing.

Inferences from Skehan-type scores concern features of discourse and tasks, while inferences from Bachman-Palmer type scores concern categories of language knowledge and degrees of mastery.

The data-based approaches to construct characterisation discussed in this chapter are summarised in Table 3. They are presented in terms of the construct studied, the purpose and approach of the study, and the implications of the findings for test development. Since each of the empirical approaches is related to a particular test or test-related research question, the aim in the studies was to specify as clearly as possible what the scores from that particular instrument mean. The implications of the results are relevant for all test developers, but especially for the planning of tests which are intended to provide detailed information on the meaning of the scores.

Chalhoub-Deville (1995, 1997) and Upshur and Turner (1998), who worked with scores-as-numbers, pointed out that the composition of the numbers in terms of components of ability which they reflect was task-specific, and the authors warned against broad generalizations. Freedle and Kostin (1993a, 1993b) also sought to show which features influenced the scores on their test, but based on analysis of test items. In all of these cases, the score is a single number which users use. The researchers investigated which concepts can or should be used to explain the scores and the differences between the score values. Chalhoub-Deville and Upshur and Turner stressed that the content information of the score varies from task to task. Chalhoub-Deville also pointed out that scores from different rater groups are influenced by different mixtures of constructs.

If strong polarisation between theoretical and psychometric construct dimensions is desired, it can be asked whether it matters that the conceptual score explanation varies between tasks or rater groups. The answer depends on the use of the scores and the values of the test developers and the score users. If the user primarily needs to know "how much" of the test's construct each of the examinees has and is happy to accept a very generic descriptor for the test construct such as "proficiency in speaking", the dependability of the numerical score is the most important criterion and its conceptual composition is secondary. If the user primarily needs to know "what" each examinee "knows" and what they do not know, the theoretical definition is important. In practice, score use rarely represents either of the extremes, which means that both dependability of scores and comprehensiveness of verbal explanations need to be ensured during test development.

Table 3. Data-based approaches to construct characterisation in language testing			
Researcher	**Construct studied**	**Purpose and approach**	**Implications for test development**
Chalhoub-Deville 1997	perceived proficiency in speaking	Identification of features which influence naïve raters' proficiency ratings. Asked groups of raters to rate samples of speech, analysed ratings, derived components.	If diagnostic information on scores is desired, this kind of method applied to a relevant group of judges would provide empirical data on specific constructs assessed.
Upshur and Turner 1999	speaking; through two task-related speaking scales	Creation of task-specific assessment scales which build on successive yes/no questions. Used sample materials to develop questions. Raters were active participants in scale construction.	If yes/no scales are required, this is the empirical method to derive them. Specificity of scale would merit further study: what if a more generic scale were used? What about score usefulness?
Fulcher 1996	fluency	Creation of a data-based rating scale for fluency. Transcribed performances, analysed examinee speech and related rater interpretations of examinee ability, constructed scale descriptors from relevant features, trialled scale.	If test developers can afford the time to develop assessment scales in this data-based way and study the functionality of the scales, the empirical backing for scores is potentially solid. Use of descriptors in feedback could be investigated.
Freedle and Kostin 1993	TOEFL reading	Explanation of item difficulty in the TOEFL reading test. Studied characteristics of texts and items, showed that difficulty was related to features of texts, not items alone.	If analysis of test characteristics is required, this method indicates which features are relevant. Suitable evidence against suspicions of construct-irrelevant variance.
Mossenson *et al;* McQueen; discussed in McNamara 1996	reading skills	Creation of maps of ability assessed in tests. Analysed content of items, inferred abilities which the items require. Created maps of item bank contents to characterise test systems, created maps individual response patterns to indicate person abilities.	This approach can be applied if a reliable basis for determining item difficulty exists. Provides diagnostic information on score meanings. Requires that inferences from task requirements to examinee abilities are reliable and that abilities cluster on a latent ability scale.
Buck and Tatsuoka 1998	listening skills	Mapping of learner attributes based on their performance on a set of tasks. Analysed tasks, derived ability attributes, analysed response patterns with rule space methodology.	If multiple skills required by items can be identified reliably, detailed maps of learner attributes can be created. Rule space methodology is statistically demanding.

Fulcher (1996b) investigated a construct only slightly less elusive than proficiency, namely fluency. Rather than studying how impressions of fluency vary, however, he concentrated on making the construct tangible enough to allow consistent assessment in a concrete assessment context. To do this, he developed a rating scale for fluency based on data from test discourse and an assessment process. His results from the trialling of the scale showed that if an assessment scale is created in such a data-based way, the assessor ratings are remarkably uniform. The method promotes close association between assessment scales and test discourse, and detailed characterisation of constructs in assessment scales. The descriptive detail also allows scale validation in relation to the testing process.

All of the data-based approaches combined data from a test with construct information from existing research to achieve an understanding of the meaning of the scores. In all cases, the data was related to test scores, and most analyses connected score categories with features of examinee performance or with task characteristics. Only McNamara's (1996) examples of item-ability maps and Buck and Tatsuoka's (1998) method of analysing learner attributes included a chain of analysis and inference from scores through features of items to categories of examinee ability. In McNamara's version, single abilities were associated with individual items, while in Buck and Tatsuoka's approach, items were characterised by a set of attributes which they require. While both authors caution that the methodology was experimental, the detail of the results is very promising when it comes to clarifying the nature of the skills assessed in a test. Both of these approaches were motivated by the wish to explain scores in detail; the other empirical approaches discussed enable this but do not necessitate it. From the point of view of score interpretation, this is what detailed definitions of the test construct can provide.

To sum up the reply to the first question presented in the beginning of the chapter, the nature of the constructs in the different approaches discussed in this chapter is either componential or procedural. The range of generalization recommended varies from strict task-specificity to rather general components which can be relevant to a large range of language use situations. Interestingly, the amount of information which can potentially be provided to examinees or other score users about the meaning of the scores does not vary on the same dimension. Fulcher's approach to developing assessment scales provided detailed descriptors while he applied the scale on learner performances on different tasks. Upshur and Turner's version of scales was highly task-specific, but all they provided for users was a number score. Thus, if the developers of a test found it

important to be able to provide detailed feedback, they would follow methodologies which have been shown to provide detailed description of the construct assessed.

As for the second question, the theorists and researchers whose work has been summarised in this chapter see the theoretical construct characterisation to be related to test development, validation, task content, the way assessment scales are defined, test implementation, and score use. In other words, the construct that a test implements is an integral element in everything about the test. The theoretical approaches provide a theoretical basis and propose ways in which tests can be related to it. The empirical approaches start from the contention that the realised test construct cannot be known without empirical investigation. A combination of the approaches enables the characterisation of a test's intended construct and the investigation of the realised construct once a test is implemented. This is probably what theorists and practitioners alike would recommend in an ideal world. However, in the practical world in which test developers need to deliver working tests to a tight schedule, detailed construct investigations may be considered extras which, while important, are nevertheless less important than meeting a schedule and proving that the test works as a measurement instrument. The question is thus both how a test's construct is characterised, and what the status of the construct characterisation in the whole venture is.

None of the approaches discussed holds an unequivocal status of importance over all others. Many testing boards are probably used to referring to theoretical models, especially componential ones, when characterising their tests. They are also used to the idea that scores need to have a label to which their meaning can be tied, be it a broad one such as "foreign language proficiency", a skill such as "reading" or "speaking", or a feature of language ability such as "fluency" or "ability to understand specific details". Furthermore, the development of a test unavoidably makes test developers aware of the features of the test and its implementation that influence the construct which is invoked in the process of assessment.

However, what this leads to in terms of how a testing board characterises the construct assessed in their test is not at all clear. Testing boards can, for instance, decide to follow a model from an existing test, such as the Australian Second Language Proficiency Interview (ASLPR) following the Interagency Language Roundtable approach to develop their oral test and assessment scale and not questioning the construct assessed. They can also relate the test to some theories, such as Hasselgren (1998),

who referred to a number of different componential models when creating an accommodated componential construct framework for her test and creating tasks and assessment scales. Analysing how the test worked as a measurement instrument, Hasselgren found that the assessors were not able to use the fluency assessment scale consistently. This led her to study examinee discourse to improve the scale. Such an interactive approach is driven by the need to make the test work, and construct characterisation is used in the process to help the developers achieve it. The construct characterisation could also conceivably be the guideline for the organisation of the whole process of test development and validation. This appears to be the assumption in existing frameworks of test development and the current state of validity theory, as was discussed in Chapters 2 and 3. In Part Two of the thesis, I will investigate reported cases of test development and validation to show how this has worked in practice.

Part Two
Reports of Practice

5 TEST DEVELOPMENT AND VALIDATION PRACTICE: THE CASE STUDY FRAMEWORK

In this chapter, I will present the framework of analysis that I will apply in the next three chapters to reported cases of test development and validation practice. The framework builds on the theories of test development, validation, and construct definition discussed in Part One of the thesis.

I will begin with a discussion of the research method, which is one variant of case study. I will explain how the principles of case study suit my research problem and present a detailed definition of the object of study. This is a summary model, based on the frameworks discussed in Part One, where test development and validation are seen as parallel processes. I will then discuss the design of the study and the selection of the cases for analysis. Finally, I will present the research questions, the procedures of analysis, and an overview of the organisation which I will follow in the case reports.

5.1 Reasons for using multiple case study

Looking at test development and validation in parallel from the point of view of test developers, the present study takes a new perspective on existing research in language testing. The case study method offers a structured research strategy for this undertaking, particularly as it supports an organized combination of theory and practice to analyse processes. A procedural view of test development and validation on the basis of existing theories and textbooks was developed in Part One of the thesis. I argued that the theoretical construct definition was one of the key quality considerations in the development and validation work. In Part Two, I will use the outcomes of the first part to analyse existing practice in language testing. The method I will use is multiple case study.

The technically distinctive features of case studies, according to Yin (1994:12-13), begin with the scope of the investigation. Case studies are empirical inquiries into contemporary phenomena within their real-life contexts, especially when the boundaries between the phenomenon and context are not clearly evident. Case studies deal with situations where experiments would be difficult to conduct because there are many more variables of interest than data points. They rely on multiple sources of evidence and they often build on theoretical propositions, which guide data collection and analysis. Yin (1994:14) also points out that case studies can

be based on any mix of quantitative and qualitative evidence, and they need not always include detailed, direct observations as sources of evidence. Typical case study topics involve decisions, individuals, organizations, processes, programs, neighbourhoods, institutions, or events.

The present study analyses test development and validation. There are indeed more variables of interest than data points in these processes, but an experiment would not be a viable research strategy in any case, because the processes are so complex and integrated into the fabric of society that their analysis in controlled designs such as those required by experimental research would not be possible. The processes are composed of multiple strands of activity which provide rich material for analysis.

While an experiment may not be a feasible research strategy, a detailed ethnographic study of a single test development process might be. A few studies of this kind have been conducted. O'Loughlin (1997) observed test design meetings as a part of a case study which investigated the comparability of a tape-mediated and a face-to-face test of speaking where the two tests were intended to be interchangeable. Lewkowicz (1997) similarly recorded the meetings of an examination board, although not in an attempt to analyse the development process but to investigate whether and how authenticity was used as a quality criterion for test tasks. Peirce (1992, 1994) recorded the process of development of a reading test form from an initial draft to a test that could be implemented, and reflected on the considerations that guided the development activity. It is notable that, firstly, each of the studies only concerned a small part of the test development process, and secondly, that these researchers were, or had recently been, members of the test development team whose work they studied, hence their access to development meetings and different versions of draft tasks. Rather than repeat these studies to deal with another test with which I was involved, perhaps covering a more extended period of time, I chose to concentrate on the range of possible practices, a focus which is one step of abstraction removed from a thorough ethnography. In order to maintain some procedural element in the analysis, however, one of my criteria for the choice of the tests to analyse was that at least one participant report on their development should be available.

A broad-range alternative to investigate language test developers' work would be to conduct a survey using questionnaires and interviews. This was done in a project which is reported in Alderson and Buck (1993) and Alderson, Clapham and Wall (1995). The survey approach affords an overview of existing practices, but it does not allow any in-depth analysis. If all of the cases included in a survey were analysed comprehensively

including detailed document analysis, the project would be quite extensive and probably no longer have the characteristics of a survey. Moreover, not all test developers document, or at least publish, their test development processes in detail (Alderson and Buck 1993:21). Detailed material on the development process and products is nevertheless necessary to analyse the nature of the process, and this was another criterion that I used to select the cases for the present study.

The present study strikes a compromise between depth and breadth. I was interested in the role of the theoretical construct definition in different test development and validation practices, so I decided to analyse development-related studies and reports. Because I believe that the participant perspective is very important in the analysis of an examination board's activities, I used the availability of at least one participant account of the development process as a criterion when selecting the cases. Furthermore, I felt it was necessary to include only cases on which several publications were available, so that the many different activities involved in test development and validation could be covered in the analysis. To avoid an issue with propriety, I decided to use only published documents.

The last point above, the decision to use published documents only, brings with it a host of caveats and limitations. Firstly, it means that the present study concerns reported rather than actual practice. Although the material includes published participant reports, these are not the same as direct observation. Rather, they are stylized accounts of activities as mediated by the report writer and possibly also a publicity board or other administrative body that governs the examination. Secondly, it means that there is an image-building aspect to the published reports. It is possible, maybe even likely, that the activities on which studies are published are presented as favourably as possible, especially since the examinations that I investigate are commercial entities. Development-related research is a key part of their competitive edge on the market. For the same reason, studies that are less than favourable for the examination may not be published, but they may nevertheless be used internally in developing the test. Other studies may not be published because the data that underlies them is not considered solid enough, or because the topics are not considered important objects of study. Some studies may be published because the examination bodies think that this is what the audience wants, and they may present rationalisations of changes that had been decided on for political reasons long before the theoretical explanations began to be sought.

Spolsky (1990, 1995) provides a carefully researched and persuasively argued perspective into the politico-commercial side of

examination development. His thesis is that institutional and social forces are more powerful explanations for practical development than theoretical evolution (Spolsky 1995:2). According to his analysis, changes in institutionalised tests such as the TOEFL are motivated by technical innovations or strong consumer demand, ie. money and power, and the threat of losing them, rather than developments in theory (Spolsky 1995:318). One of the implications of Spolsky's study for the present thesis is that it does not cover all the concerns in real-life examination publishing. Another more practical one is that the reports that I study are likely to reflect "desirable practice" because they essentially represent one voice, that of the examination board as it wants to present itself to researchers and to test users.

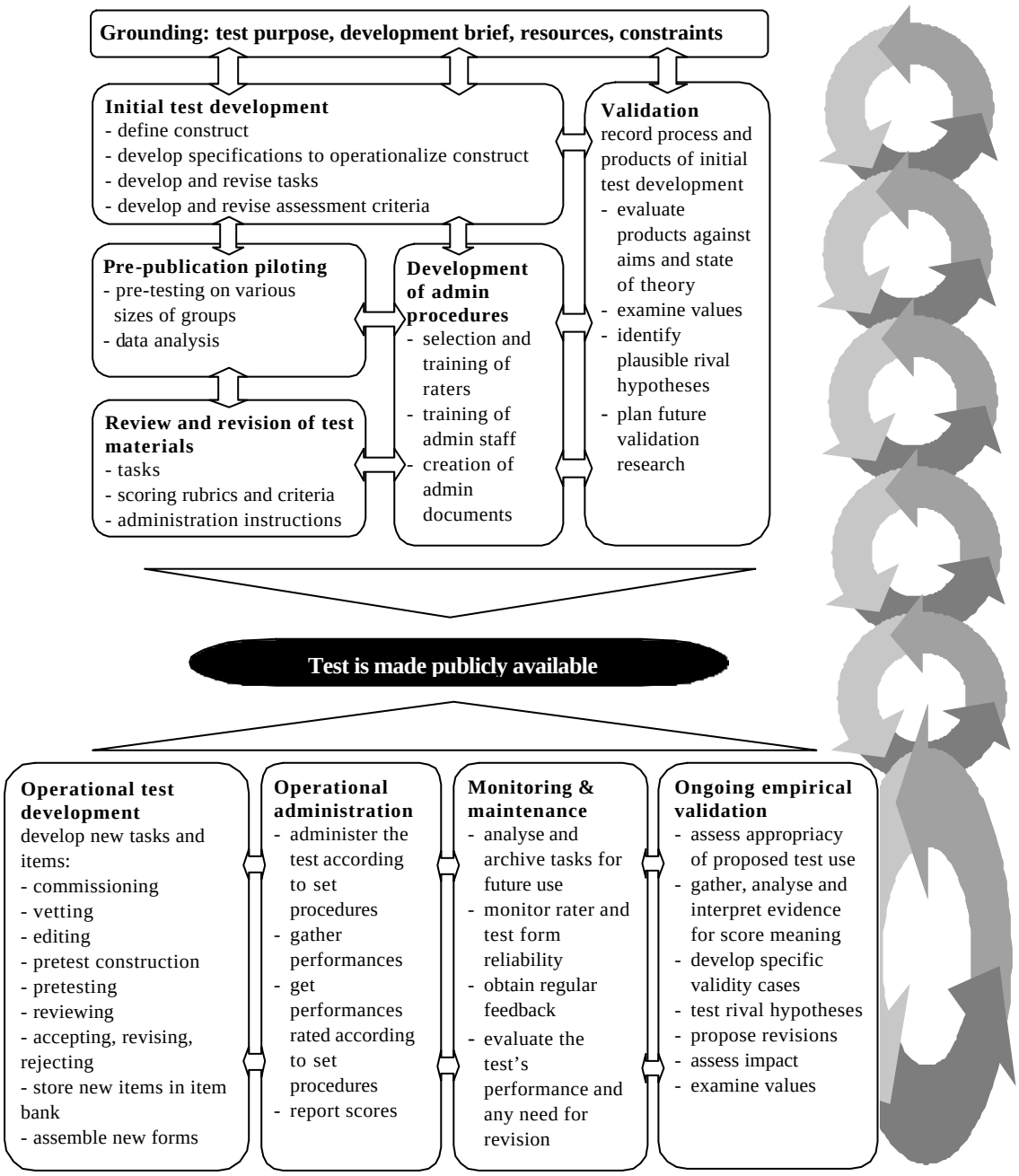
Regardless of the caveats and the sobering reminder of the narrowness of the theoretical perspective, I will concentrate on the content and construct agendas in test development and validation in the present thesis. I believe that the studies published on a test reflect some of the actual development activities and also the developers' beliefs on acceptable practice. Moreover, I believe that the reports reflect the values of the test developers and/or the examination boards in terms of what they consider important about the nature of the construct and the purposes of measurement. This is what I study in the present thesis.

To make case studies rigorous and accountable, Yin (1994:63) recommends the use of case study protocols. The protocol contains the case study questions and the procedures to be followed when they are asked. According to Yin, the protocol is an essential research tool in multiple-case designs, as it helps ensure cross-case consistency in the analysis, thus supporting the reliability of the study. It should contain an introduction to the project, a summary of the field procedures, the case study questions, and a guide for the case study report (Yin 1994:64-70). Yin notes that the questions in the case study protocol concern the *individual case*, not cross-case comparisons, and they are questions *to the case study researcher*, not directly to informants. Similarly, the guidelines for how to report a case study are guidelines for the researcher. A plan for the report outline before the actual analyses are conducted is important, according to Yin (1994:73-74), because case studies do not have clear, uniform guidelines for reporting similar to those for reporting experiments, for example. In this chapter, I will present the case study protocol which I will follow in the next three chapters of the study.

5.2 The object of analysis: the test development and validation process

A summary model of language test development and validation, based on the theories and frameworks discussed in the previous three chapters, is presented in Figure 3. This model describes the object of analysis in the case studies below. When analysing the reports of test development and validation, I will use this model as a framework for structuring the analysis.

Figure 3. The test development and validation process.



The activities in Figure 3 are described from the point of view of test developers, and they are grouped together under conceptual headings. The model illustrates the relationship between test development and validation as it is described in current theoretical frameworks. The activities and processes are placed on a rough time line which begins at the top of the figure and runs downwards. The only point on the timeline that can be clearly pinned down, however, is the publication of the test. The activities on either side of this dividing line are potentially simultaneous and interdependent. The circles on the right of Figure 3 denote this cyclical and iterative nature of the development and validation processes.

The activities related to test development and validation run in parallel in the model. My contention, based on the theoretical reviews in Chapters 2 and 3, is that this is the way they are currently characterised in theoretical writing, even though validation literature has not emphasized the procedural nature of test-related validation work. The practical real-world grounding for both activities is provided by the concrete situation in which the test is being developed: the purpose of the test, the task which the developers are given, the resources available, and the practical constraints in which the development and later operation of the test are to take place.

The test development activities prior to publication encompass the development of all the texts and procedures needed in the operational use of the test, the training of necessary personnel to administer and assess the test, and the trial runs with the tasks and administration and assessment procedures. The decisions taken on what should be measured and how the quality of measurement is guaranteed reflect the values of the test developers and/or the political decision making body that is responsible for overall decisions. The activities are intended to optimise the test and its implementation and to gather data on the quality of the test. The validation activities prior to the publication of the test comprise the recording of the process and products of the test development as well as the initial investigations into the construct to be measured. They also include the design of validation studies to be conducted during the operational stage of the test.

At publication, many of the administrative procedures are set and the operational administration begins. At the same time, if not before, the construct intended to be measured is also consolidated, since the expectation is that different versions of the test which have been developed according to the same blueprints and administered according to the same procedures implement the same construct. The nature of test development and validation changes slightly after the publication of the test because the

test is operational, and the system and its scores are used in society. During the operational stage, in addition to the intended construct, there is the implemented construct, whose indicators include the actual testing procedures and especially the scores which are produced by the test and used in society. Furthermore, there are the users' interpretations of the construct, which are realized in the uses to which they put the scores and the perceptions that they have of the assessment system in general. During both initial and operational development, the validation studies actually conducted reflect the values of the test developers with respect to the nature of the construct and the procedures needed to guarantee the quality of the system.

Although test validation and use involve several groups of actors, my focus in the present thesis is on test developers. While I consider post-publication activities of test development and validation a relevant object of study, score users' interpretations of the construct provide relevant material for the present thesis only to the extent that test developers make use of them in their development and validation activities.

5.3 Rationale for the case study design

The design of the multiple case study in the present thesis is based on several considerations, all of them tied in one way or another to the definition of the object of study in Figure 3. I wanted to analyse the entire process of test development and validation, and thus it was necessary to ensure that both initial and operational test development would be covered by the cases. This would also enable me to analyse similarities and differences between initial and operational test development and validation. Since my basis of comparison is recommendations from theory, it made sense to select tests that would be likely to implement theoretical recommendations as carefully as possible. This would be the case with high-stakes tests, because these, if any, are likely to be held publicly accountable of their practices. And since theorists of test development and validation agree that purpose is the most important guideline for test development and validation, it was necessary that the tests I analyse serve the same purpose.

The design of the case study was further influenced by my interest in the role of the theoretical construct definition in test development and validation. I believe that from the perspective of theoretical design, the construct definition guides the questions asked and the procedures followed when test development decisions are made. Since theoretical rationale rather

than political decision-making practice is what I study in the present thesis, this is a central object in my study.

My belief in the significance of the theoretical construct definition is what Hamel et al. (1993:44) would call an *initial theory* of the present thesis. They define initial theory as "the initial idea that a researcher had of the perceived … issue or phenomenon" (Hamel et al. 1993:44). The initial theory influences the way that the object of study is defined, and it must be related to existing theories in the field being investigated. Since Hamel discusses case studies in sociology, he anchors the concept of initial theory to social issues and sociological frameworks. In the case of the present thesis, my belief or initial theory is related to language testing, educational measurement and applied linguistics. Its grounding in this research was presented in Part One of the thesis.

Although the concept of *initial theory* is closely related to that of the *working hypothesis* in empirical research, I do not want to use the empirical research terminology. I will call my belief a belief or an initial theory, and consider it to lead to *expectations* about what the case analysis will show. The use of this terminology is by no means intended to downplay the defensibility or rigour of the case study as a research approach. I simply want to make it clear that the conceptual world of the present study is parallel but not equal to that of experimental research. Whereas working hypotheses are often pursued in experimental research to develop specific hypotheses that can be tested in future research, there is no such wish in the present study. Neither is the current design an ill-defined experiment. The design is intended to throw light on the complex processes of language test development and validation. The results are intended to specify the theoretical frameworks that the language testing research community uses for test development and validation so that they would help practical test developers develop quality into their tests.

For the purposes of the case study, I divide test development rationales into three possible categories according to the nature of the construct definition in them. In the first category, the theoretical construct definition is brief whereas the psychometric construct definition is detailed. In the second, the theoretical construct definition is extended whereas the psychometric construct definition is less extensive than in the first category. In the third, both the theoretical and the psychometric construct definitions are extended. I have chosen one case of reported practice to represent each of these three conditions.

Since the design in principle follows a two by two matrix where one dimension distinguishes between theoretical and psychometric construct

definitions and the other between brief and extended definitions, I must explain why I do not discuss four cases. The missing case is one where neither the theoretical nor the psychometric construct definition are extended. I do not discuss it because I do not believe this would be considered acceptable practice and because I believe it would not be likely that the development of such a test would be reported and published if it did exist. The test development cases that have been reported in the language testing literature do not represent perfection, of course, but those that have been reported through extensive reports do represent serious effort to meet quality requirements.

My initial theory and the case design lead to two main expectations. Both build on the assumption that emphasis on numerical construct definitions leads to a reliance on psychometric arguments to support the quality of the test while emphasis on the verbal construct definition leads to a reliance on theoretical quality criteria that concern test development rationales, validation studies, and the relationship between test development and validation. The first expectation is that the reports of test development will clearly reflect differences between the cases. The discussion in Chapter 2 showed that test development is a flexible and multi-dimensional process, and I expect that the type of quality standards used in a case will guide the questions asked, the materials investigated and the rationales presented for different test development decisions. The second expectation is that the association between the construct definition and the validation rationale will not be equally clear. As was discussed in Chapter 3, there are very few instructions and examples about how to use the theoretical construct definition in validation, so it is possible that all the cases show a concentration on psychometrically motivated validation designs.

If the initial theory is entirely wrong, the reports of test development and validation will only differ in terms of how they characterise the construct assessed, which was my basis for categorising the cases in the first place. If the analysis shows that there are clear differences in the reports but that the cases differ in many terms so that it cannot be said how far the nature of the construct definition can explain them, the initial theory was too narrow and a more detailed proposal can be developed. The initial theory cannot be confirmed, but support for it would be afforded if the expected differences in test development and validation procedures were revealed in the analysis.

The reason why I focus on the role and nature of the construct definition in test development and validation is my perception that the advice that theory provides is patchy and somewhat contradictory. Current validity

theory, in my interpretation, assumes that test development is guided by a theoretical construct definition. The centrality of the construct is not questioned; the discussion concerns how far into construct theory the test developers' work should go. Somewhat in contrast, I suggest that current test development frameworks propose a distributed focus for development and validation, so that the construct is one criterion, but there are also other important concerns. Bachman and Palmer (1996), for instance, propose usefulness, while Alderson et al. (1995) and Weir (1993) propose validity, reliability, and practicality. Both validation and test development theory eschew the reporting of psychometric information only, with no theoretical elaboration of the construct. However, there are clear technical and methodological guidelines for how to provide psychometric information for tests. In contrast, guidelines for how to describe, discuss, and investigate the intended or realised construct on a theoretical/descriptive level are at a much earlier stage of development. If the theoretical construct definition is indeed as important as current validity theory states it is, test developers need advice and examples of its actual role in test development work.

In summary, the purpose of the multiple case study in the present thesis is to study the nature of reported practice in test development and validation, and especially to investigate the role of the verbal construct definition in it. The results show examples of different practices and enable comparison between theory and reported practice.

5.4 Selection of cases

The criteria I used to select the cases for the present study were:

- availability of at least one participant report on the test development process
- availability of detailed material on test development and validation in the form of several published reports
- serving of the same test purpose
- coverage of initial and operational test development and validation across cases
- representation of one of three categories with respect to the nature of the construct definition in the test:
  - theoretical definition brief, psychometric definition extended
  - theoretical definition extended, psychometric definition less so
  - extended theoretical and psychometric construct definition

There were a range of publications which met the first criterion, availability of at least one participant report. These included case studies, PhD theses, and brief accounts evidently mostly intended for test users. With the introduction of the second criterion, the number of potential cases dropped drastically, leaving four. With the introduction of the third criterion, the Australian **access:** test was dropped because it was used as a selection criterion in immigation decisions rather than university entrance. Thus the criteria resulted in the selection of three cases, which together covered initial and operational test development. The first was the paper-based TOEFL Reading test. The reason why this case is delimited to the Reading section only was that this was the only section for which a participant report existed (Peirce 1992, 1994). Since a range of other studies also existed on this section, it was possible to include it in the analysis. This case concerned operational test development. The second case was the development of IELTS (eg. Alderson and Clapham (eds.) 1992). Several participant reports existed on the initial development of this test, and a range of publications were also available from its operational stage. The third case was TOEFL 2000, which is committed to a combination of theoretical and psychometric construct considerations in the development of the test (Jamieson et al. 2000). This test development effort is ongoing, and has not yet reached the prototyping stage.

5.5 Case study questions

My main research questions in the case studies of practice are:

- What are the similarities and differences between initial and operational test development and validation?
- How does the nature of the construct definition in a test influence the rationale followed and the quality criteria used in test development?
- How does the nature of the construct definition in a test influence the rationale followed, the questions asked, and the results published on validation?
- How do the examples of realised practice in language test development and validation correspond with recommendations from theory?

These questions can be answered at the level of cross-case comparison. However, as discussed earlier in this chapter, the case study

protocol should contain the questions in the form that they will be asked within each case. These will be presented below.

The case analysis proceeds in three steps. The first is an initial analysis of the nature of the documents and reports that have been published on the test. The aim is to characterise the general approach that the test developers take to their work. The second is a more detailed analysis of selected documents to investigate the processes of test development and validation and the role of the construct definition. This provides an account of the activities conducted. The third step is a summary of the values that seem to guide the development and validation decisions and activities. The detailed questions that guide this analysis are presented in Table 4.

Table 4. Case study protocol

General approach to test development and validation		
- What topics are addressed in the publications related to the test? - What materials and methods are used in the studies? - Which concerns in test development and validation do the studies address? - How is the construct investigated or operationalized in the studies?		
Test development	**Validation**	**Construct definition**

- What steps, stages, or parts do the developers identify in test development?
- How does this compare with the stages of test development in the case study framework?
  - Any areas not reported on?
  - Any additions?
- What questions and/or criteria guide the development of the test?
	- How do the test developers see validation? How do they define it?
	- What material do validation studies refer to?
	- What questions does validation work seem to be guided by?
	- What do the test developers report about validation results?
		- How do the developers describe or define the construct their test is assessing?
		- What conceptual categories or perspectives does the construct definition contain?
		- How was the definition developed and when?
		- How is the construct definition operationalized in test development and validation?
		  - verbally? numerically?

Values that guide test development and validation		
Judging by the test content and format, what aspects of language have to be tested? Judging by the assessment procedures selected, what produces quality in assessment? Judging by the studies related to the test, how do the test developers justify their development decisions?		

5.6 Materials analysed

As was indicated earlier in this chapter, the present thesis is an atypical case study in that it does not include direct observation or interviews. The

reasons for this are political and practical. Publicly available language tests are commercial products, and their proprietary nature makes it unlikely that I as a researcher would have been allowed to observe actual test development in different test development teams. Furthermore, the cases that I included in the present study come from different points in time; thus it would have been difficult to conduct interviews even afterwards, and the material from any interviews actually conducted would not have been comparable because subsequent events would have changed the perceptions of the developers of earlier tests. My material for the analysis of the three cases, instead, consists of the published reports and studies on the development and validation of the three test development efforts. Each set of material includes at least one participant report of the development activities. A closer characterisation of the materials is presented in the case reports. In addition, I will refer to current publicity material for each test for background information.

5.7 Organisation of the case reports

The organisation of the case reports in the next three chapters will follow a set pattern. Each case begins with an introduction, which characterises the setting of the development project that will be analysed in the case. The test purpose, the testing board, the time when the test was introduced, and its current width of use will be presented. The boundaries of the case will be defined in terms of time and the stages of development. The test instrument will be characterised as to its sections and tasks, and a brief overview will be given of the changes made in these during the period that I analyse in the case. The roles and responsibilities of the test developers and the administrative body will be described according to what is said about them in the test development reports and publicity material. Finally, before the actual analysis, a summary will be given of the conditions and constraints in the test development and validation work as they are characterised by the test developers themselves.

The case analysis begins with the first step described in section 5.5, an analysis of the nature of the reports that have been published about the development of the test. This is followed by a descriptive report on the test development and validation activities in each case. The description is based on the studies published, and the presentation is organised according to the categories of the summary model presented in Figure 3 in the present chapter. I will separate test development and validation in the case reports as far as is justifiable on the basis of the source documents. Each case

report concludes with a summary where the questions of the case study protocol, presented in Table 4, are answered.

In Chapter 9, I will address the overall case study questions and discuss cross-case comparisons. I will discuss the degree to which the expectations discussed in the present chapter were borne out in the case analysis, consider alternative and additional perspectives that had not been included in the original case study design, and draw implications for further study.

6 BRIEF THEORETICAL DEFINITION OF CONSTRUCT: PAPER-BASED TOEFL READING

6.1 Introduction to the TOEFL Reading case

The Test of English as a Foreign Language (TOEFL) evaluates the English proficiency of people whose native language is not English (Educational Testing Service (ETS) 1999a:3). The test includes items in listening, structure, reading, and writing; the test of speaking is a separate test for which participants can register if they want. TOEFL is developed by the Educational Testing Service (ETS). It is administered in more than 1,275 test centres in 180 countries around the world (ETS 1999a:4). In the operational year 1997-1998, some 930,000 people registered for the test. In 1998-1999, during the gradual introduction of the computer-based version, more than 300,000 people registered for the computer-based test and more than 380,000 for the paper-based test (ETS 2000a:4, ETS 2000b:4). The paper-based tests are arranged on set dates on a monthly basis; the computer-based tests can be administered at any time that is convenient for the test taker.

The TOEFL test was developed in co-operation between more than 30 organizations, and it was first introduced in 1963 (ETS 1997:7). Since then, the test has undergone two revisions. The first was made in 1976, when the current three-section version was developed out of a previous test structure with five sections (ETS 1997:11). The present, ongoing revision programme is called TOEFL 2000. Changes related to it are implemented gradually. The first changes were introduced in 1995, when some changes were made in the basic TOEFL (see below) the Test of Spoken English was revised (ETS 1999c:3). The next step was the introduction of test delivery by computer. This was started in 1998 and is ongoing (http://www.toefl.org/develop.html). Both computer-based and supplemental paper-based test sessions are currently arranged around the world depending on availability of computer-based testing facilities.

6.1.1 Boundaries of the TOEFL Reading case

The present case focuses on the paper-based TOEFL Reading section from the first revision to the present day. The Reading section was selected because the literature on this section fulfilled the criterion that a participant report was available on the actual process of test development. Since this

report concerned the paper-based version of the test, this is what will be analysed in the present case. The 1976 revision is not included in the analysis because I did not have enough material on the process of test development activities then. In terms of time, the present case nevertheless spans nearly twenty-five years, starting from 1976 and coming up to the present day. In terms of stages of test development, it focuses on operational test development. Thus, in relation to the model of test development that I discussed in Chapter 5 (see Figure 3), the TOEFL Reading case concerns activities that happen *after* the publication of the test, below the black line in Figure 3. During the operational period that I investigate, the TOEFL Reading section underwent one revision.

The TOEFL Reading case represents the category where the theoretical construct definition is brief and the psychometric construct definition is extensive. Motivation for this categorisation is given eg. by Spolsky (1995:3), who describes the multiple-choice TOEFL test as a prime example of tests which "place their highest value on technical reliability, efficiency, and commercial viability". The TOEFL board's position on the nature and status of the theoretical construct definition has changed with the TOEFL 2000 programme, this will be discussed further in Chapter 8.

6.1.2 Format of the TOEFL Reading section

The Reading items in the paper-based TOEFL test belong to what is often referred to in ETS publications as Section 3 (eg. Duran et al. 1987:56, ETS 1999a:15). The present case covers two formats for Section 3, one where the section had two parts, and one where the parts were integrated. The change was introduced operationally in April 1995.

Before the change, Section 3 was called Vocabulary and Reading Comprehension. It contained 30 vocabulary items and 30 reading items that counted towards candidate scores, and some items that were being pretested and therefore did not contribute to candidate scores (Peirce 1992:668). The test format was four-option multiple choice throughout the section. The examinees were asked to choose the best of the four alternatives. The vocabulary items were discrete, and tested the candidates' knowledge of words in the context of one sentence. The reading items were in what Peirce calls "sets", or what are technically called testlets. Wainer and Kiely (1987:190) define a testlet as "a group of items related to a single content area that is developed as a unit." The TOEFL reading items fulfill this criterion because a group of them focus on the same text passage. In the pre-1995 TOEFL reading test, there were usually five reading passages with approximately six items on each of them (Peirce 1992:668).

The revision in 1995 changed the format of TOEFL Section 3 by eliminating the separate vocabulary subpart (Schedl, Gordon, Carey, and Tang 1996:2). The traditional four-option multiple choice continued to be the item type used. The revised test included approximately five passages with approximately ten items each (inference from Wainer and Lukhele 1997:7). Any test form which is actually administered may include a sixth passage, however, as pretest items can be inserted in operational test forms (Schedl et al. 1996:10). As before, the pretest items do not count towards the examinees' scores. The approximately five-passage, approximately 50-item multiple choice format is still used in the paper-based administrations of the TOEFL test (ETS 1999a:15). The computer-based version includes new item types (see eg. ETS 2000b for a description and examples). The analysis in the present case is confined to the paper-based version of the test.

According to ETS (1997:12), Section 3 measures "the ability to read and understand short passages that are similar in topic and style to those that students are likely to encounter in North American colleges and universities." The items may focus on factual information presented in the passages, but examinees can also be asked to make inferences or recognise analogies. (ETS 1997:12.)

The TOEFL score reports record an overall score and section scores for listening, structure and written expression, and reading (ETS 1999a:34). A separate score is reported for the Test of Written English if the candidate has taken it (ETS 1999a:35). In the scoring of the paper-based TOEFL, each correct answer counts equally towards the score for that section. The computer-based score report follows the same general format, but because the test and the scoring procedures are slightly different, different reporting scales are used (ETS 1999b:17). ETS provides concordance tables for comparing the two sets of scores (eg. ETS 1998: Appendix A).

6.1.3 Developers of the TOEFL Reading test

The development of the TOEFL test involves both those who write and revise the actual test items and the organization at ETS that develops the whole examination, sets policy, and guides research. Both are important as sources of information on how the TOEFL test is developed, and the reason that I selected the Reading section for analysis in the present chapter was that publications were available on both perspectives.

Draft tasks for the TOEFL test are written by language specialists outside ETS who receive training and then start to write draft tasks

according to detailed test specifications (ETS 1997:12, Peirce 1992:669). The test specifications are developed, revised and approved by the TOEFL Committee of Examiners (ETS 1997:12). The specialists send the completed drafts to the ETS test development department, where a member of the test development team takes the responsibility of converting them into a publishable pretest set (Peirce 1992:669). This involves iterative cycles of comments from colleagues and revisions to the items (to be discussed in more detail later in this chapter). The items are then pretested and final revisions are made in the light of empirical data (Peirce 1992:669).

The pretesting and construction of operational test forms involves a number of different departments at ETS. In her report of the case study on TOEFL Reading, Peirce (1992:677) mentions the Test Development department where she worked and the Statistical Analysis department. Given the confidentiality of the examination and the large number of tests administered each month, there must also be a printing and test production unit, which Peirce mentions indirectly when she talks about galleys for the test being returned to the Test Development department for a final review before it is published (1992:673).

Educational Testing Service develops the TOEFL under the direction of an administrative body that is called the TOEFL Board, formerly known as the TOEFL Policy Council (as reported on the TOEFL website, http://www.toefl.org/edgovbod.html). The board has fifteen members including official representatives from other ETS boards, representatives of TOEFL score users such as colleges and universities, and specialists in the field of English as a foreign or second language. It also has five standing committees. The most important of these for the present case is the TOEFL Committee of Examiners, which sets guidelines for the TOEFL research program.

According to the TOEFL website, the Committee of Examiners has ten members, some of whom come from the ETS and some from the academic community. However, research on TOEFL data is still mostly conducted by ETS staff (ETS 1999c:3). A core area of research identified by the Committee of Examiners and TOEFL staff is “enhancing our understanding of the meaning of TOEFL scores” (ETS 1999c:1). The results are published in three series of reports: TOEFL Research Reports, TOEFL Technical Reports, and TOEFL 2000 monographs. The reports in these series are a key source of material on the practices followed in the development of the TOEFL test.

6.1.4 Test development brief: conditions and constraints

Since the present case concerns operational test development, the phrase "development brief" essentially refers to the test specifications and the policies that govern the overall development of the examination. The specifications for the TOEFL test, similarly to other publicly available tests, are confidential. As discussed in Chapter 2, it is likely that they include rules about features such as the number of items on the test, item format, item focus, and test time. The published features of the specifications were discussed at the beginning of this chapter: the test format is four-option multiple choice, the test tests ability to read and understand short passages of academic-type texts, and the items focus on information given in the text, inferences, and recognition of analogies (ETS 1997:12).

The policies that guide the development of the TOEFL test are likely to focus on the required measurement features of the test forms such as difficulty and discrimination. Similarly to the specifications, these are not published in detail, but some information about them is available. According to the *TOEFL Test and Score Manual* (ETS 1997:12), the measurement qualities of the test are monitored by using pretest information when operational tests are constructed and by routine analyses of data from operational administrations. The main method of guaranteeing score comparability from administration to administration is score equating (ETS 1997:22-23, 29-33). In brief, the developers implement this by statistically linking the raw scores that examinees gain on each test administration to a common reference point so that they can be reported on a common scale. Some of the studies on the TOEFL test examine the methods used to meet these conditions and constraints. This research will be discussed below.

6.2 Nature and focus of studies published on the TOEFL Reading section

In this section, I will report on a preliminary analysis of the kinds of questions pursued in the existing literature about the development and validation of the TOEFL Reading test. I will mostly discuss studies published by the ETS, but a few studies published externally to the testing board will also be included. I will discuss the studies in three groups: TOEFL Research Reports, TOEFL Technical Reports, and other studies relevant to the development and validation of the Reading section.

The ETS magazine *The Researcher* (ETS 1999c:2) groups the topics of the TOEFL Research and Technical reports and TOEFL Monographs into eight areas of inquiry as indicated in Table 5. The areas are test

validation, test information, examinee performance, test use, test construction, test implementation, reliability, and applied technology – each with further sub-divisions. This broad research agenda covers test development and validation from several perspectives, including the test developers' work (test construction, implementation, and use); the measurement qualities of the test instrument (validity and reliability); and the interpretation of examinee performance both in terms of content and processing and in terms of test scores.

The Researcher (ETS 1999c:2) helpfully cross-references each of the published TOEFL Research Reports, Technical Reports, and Monographs by the areas of inquiry that they address and by the sections of the TOEFL test that they investigate. Any report can concern more than one inquiry area and more than one TOEFL section, and they very often do. For instance, Research Report 1, *The performance of native speakers of English on the Test of English as a Foreign Language* (Clark 1977) addresses test validation (face/content validity), examinee performance (difference variables), and test use (decisions/cut scores) concerning all the three sections of the TOEFL as well as the test in general.

The table in *The Researcher* (ETS 1999c:2) helped me identify studies published by the ETS which are relevant to the Reading section. I looked up the studies and analysed their nature and approach. There were thirty Research Reports that were indicated as relevant to the Reading section and eleven Technical Reports. Tables that give detailed information about these reports can be found in Appendices 1 and 2. Appendix 1 covers the Research Report series, and Appendix 2 the Technical Report series. The tables identify the reports by their number in the series and their author and date of publication, and also specify the aims of each study, the materials and methods used, and the main findings. The last column in the tables indicates the areas of inquiry from those presented in Table 5 (ETS 1999:3). In most cases, at least two areas were identified by the publisher for each report. When a report only focused on one area, it tended to be construct validity in the Research Reports series and test equating in the Technical Report series. Taken together, the reports cover all the main categories in Table 5, and the most frequent concern addressed is validity.

A first impression arising from the list of the Research Reports that ETS (1999:3) has indicated to be relevant to the TOEFL Reading section (see Appendix 1) is that many of the studies focus on TOEFL scores (eg. Clark 1977, Pike 1979, Angelis et al. 1979, Powers 1980, to name the first four). The studies investigate how the TOEFL total scores or section scores vary and what the scores can be considered to indicate.

Another identifiable group of studies investigates *item* properties and item responses to study the more detailed nature of the construct measured in the test or to study item bias (eg. Alderman and Holland 1981, Oltman et al. 1988, Freedle and Kostin 1993b). A third group of studies investigates the alternative scenario to the score-driven investigation of what the TOEFL tests. These reports (eg. Duran et al. 1985, Stansfield (ed.) 1986, Henning and Cascallar 1992) start from a concept of communicative competence and study the extent to which the TOEFL test can be considered to measure aspects of it.

Table 5. Areas of inquiry in TOEFL research published by ETS (ETS 1999c:2)

1. test validation
   a. construct validity
   b. face/content validity
   c. concurrent validity
   d. response validity
2. test information
   a. score interpretation
   b. underlying processes
   c. diagnostic value
   d. performance description
   e. reporting/scaling
3. examinee performance
   a. difference variables
   b. language acquisition/loss
   c. sample dimensionality
   d. person fit
4. test use
   a. decisions/cut scores
   b. test/item bias
   c. socio/pedagogical impact
   d. satisfying assumptions
   e. examinee/user populations
5. test construction
   a. format rationale/selection
   b. equating
   c. item pretesting/selection
   d. component length/weight
6. test implementation
   a. testing time
   b. scoring/rating
   c. practice/sequence effects
7. test reliability
   a. internal consistency
   b. alternate forms
   c. test-retest
8. applied technology
   a. innovative formats
   b. machine test construction
   c. computer-adaptive testing
      item banking

The studies that primarily analyse TOEFL total and section scores investigate them in relation to each other, in relation to other tests, in relation to other test types which could possibly be used in a revised TOEFL test, between native speakers and non-native speakers, between different cultural and native language backgrounds, and for the same individual at different points in time. The sample sizes are usually large and the methods carefully selected to fit the purpose of the study. The purposes are related to understanding the variability in TOEFL scores and, in factor analytic studies, to the factor structure that underlies the score distributions. Because of the focus on score interpretation, these studies are related to construct validity, and this is how most of them are categorised in *The Researcher* (ETS 1999c:3). However, the logical structure of the studies

starts from a numerical score and investigates its content meaning. In other words, the starting point is the fact that there is a reliable measurement scale, and the question is what it indicates. Questions of what it does *not* indicate, ie. questions of construct representativeness, are not asked. Studies on the test's factor structure, for instance, investigate the nature and composition of the psychometric constructs that underlie the scores. All the studies point to the conclusion that the listening section can be separated from the rest of the sections for all language backgrounds and all ability levels. The studies also regularly yield 2-3 further factors, but their possible interpretation varies between different examinee groups (Hale et al. 1988:51-52). This may mean that the test's factor structure varies and the justification for reporting an overall score and three section scores is not supported by the score data, but it may also indicate genuine differences between examinees from different language backgrounds or with different levels of ability.

In addition to score-based construct questions, the studies published in the TOEFL Research Report series also investigate the quality of the test when it is used. Through score data, the fairness of the test is studied especially in terms of test or item bias. To illustrate, Alderman and Holland (1981) investigated item bias and found that nearly seven eighths of the TOEFL items were found to be sensitive to the examinees' native languages. Specialists attributed the differences to similarities between English and the native language, but were unable to predict bias on the basis of looking at the task alone, without response data. No clear conclusions for the rules of test construction were drawn. Angoff (1989) studied whether examinees tested in their native countries were disadvantaged because of American references in the text and found that they were not. The score data also enables the test developers to investigate plausible rival hypotheses to score explanations, notably speededness. Secolsky (1989) and Schedl et al. (1995), for instance, found that speededness may be a problem in the test in general and in the Reading section in particular, especially when pretest items are inserted in operational forms.

A group of score-based studies on the TOEFL (Angelis et al. 1979, Powers 1980, Alderman 1981, Wilson 1982) investigated the role of English language ability as a moderator variable in the assessment of academic aptitude. The studies were motivated by the fact that applicants to North American universities were often required to take tests of academic aptitude as well, and it was possible that the interpretation of their scores in terms of academic aptitude was questionable. The data analysed was scores on TOEFL and the aptitude tests, and non-native speaker performance on the aptitude tests was compared against native speaker performance. The main

findings were score ranges above which the interpretation of the aptitude test scores for non-native speakers became meaningful. The studies provided some evidence that the TOEFL test can be used meaningfully as a measure of language ability since it helps explain the academic aptitude scores of non-native speakers with low proficiency in English. This is important information for administrators who admit students to colleges and universities. However, the contribution of these studies to the definition of the construct assessed in the TOEFL is only that the test assesses language ability.

More detailed analyses of the constructs assessed in the TOEFL Reading section have also been published in the Research Reports series. These use data on item performance and concepts stemming from the test specifications to study the relationship between the two. Freedle and Kostin (1993b) investigated the degree to which the textual characteristics of TOEFL reading passages and reading items can explain item difficulty. This study was discussed in Chapter 4. Schedl et al. (1996) investigated whether the "reasoning" items in the Reading section measure something different from the rest of the reading items. Such investigations are important for the present study because they show how a test whose construct definition is numerically oriented makes use of verbal explanations for the construct assessed.

On the basis of the list of study purposes and materials in Appendix 1 and the discussion above, the research questions addressed in the TOEFL Research Reports related to the Reading section can be summarised in the following list:

- How do the TOEFL total and section scores vary between different examinee groups?
- How could or should the scores be reported?
- Which factors explain the scores?
- How are the TOEFL scores related to scores from other language tests or to scores from tests of academic aptitude?
- Are the TOEFL scores or items biased against identifiable groups of examinees?
- How is the TOEFL test related to models of communicative competence?
- What does the Reading section measure?

In the more detailed analysis later in this chapter, I will only analyse some of the Research Reports listed in Appendix 1. The selection criteria were that the studies should focus on the development and validation of the

Reading section rather than the whole TOEFL test and that they should focus on the test development procedures, validation procedures, or construct definition. Three whole reports were thus selected for further analysis: Freedle and Kostin's (1993b) study of item difficulty in the reading section, the Schedl et al. (1995) study on the development of the all-passage reading section, and the Schedl et al. (1996) investigation of the reasoning and non-reasoning items in the reading section. Parts of reports which provided further relevant information included Duran et al.'s (1985) analysis of test content, the sections of papers from Stansfield (ed. 1986) that discussed the reading section, Henning's (1991) study of vocabulary items embedded in reading paragraphs, and Boldt et al.'s (1992) investigation of the relationship between TOEFL scores and ACTFL ratings. Although Henning and Cascallar's (1992) preliminary study of the nature of communicative competence is interesting in the light of future developments in the TOEFL 2000 program, its contribution to the analysis of what is assessed in the Reading section proved to be so small that its analysis in the present case was not justified.

The TOEFL Technical Reports (see Appendix 2) constitute a newer series than the Research Reports; the first Technical Report was published in 1991. Judging by the aims of the studies, the reason for the word "technical" is an emphasis on the *methods* used in test construction and data analysis. That is, many of the reports have a distinct *how to* aim: how to equate test forms efficiently, how to investigate speededness, how to use the quantitative information available from TOEFL scores in meaningful score reporting. The conclusions in these studies are also related to the efficiency of the methods used in the study for the purpose of the study, such as test equating.

The data used in the Technical Reports related to the Reading section is most often different kinds of score data: total and section scores, item scores, detailed item responses (correct, incorrect, omitted, not reached), and item parameters. These are often real data, but some studies also use artificial data and compare the predictions and estimates made on the basis of artificial data or prediction algorithms with those based on real data. Three of the reports also use other data related to the items: Way et al. (1992) used information on item position in a sequence of items and information on length of time between pretesting and operational use, Chyn *et* al (1995) used content-based rules for test construction presumably stemming from the test specifications, and Boldt and Freedle (1996) used the same 75 textual characterisations of items and passages that Freedle and Kostin (1993a, 1993b) had used in their study of prediction of item

difficulty in reading comprehension. The conclusions in these studies made use of the non-score information to explain patterns of score variance.

The Technical Reports described above can be divided into five topic areas: test equation, item selection and test construction, meaningful score reporting, effect of test format on reliability, and the threat of speededness as an alternative hypothesis for explaining TOEFL scores. They are thus clearly relevant for an analysis of routine and experimental test construction mechanisms and validation procedures.

The research questions investigated in the Technical Reports series can be summarised as follows:

- How could the TOEFL test forms be equated efficiently?
- How can speededness be investigated efficiently?
- How can TOEFL test forms best be constructed?
- What skill information might be available from the TOEFL scores?
- What might explain differences between item parameters in pretesting and operational use?
- How reliable is the TOEFL test?

In the discussion below, I will refer to most of the Technical Reports where they provide information on the construction and validation of the TOEFL Reading section. More detailed analysis is due to Wainer and Lukhele's (1997) study of test reliability, Way, Carey and Golub-Smith's (1992) analysis of differences between pretest and operational item characteristics, Chyn, Tang and Way's (1995) study of automated test construction procedures, and Boldt and Freedle's (1996) study of predicting item difficulty.

Outside the TOEFL report series, there are a vast number of studies in which the TOEFL test has been used in some form. However, I will concentrate here on studies which give information about the development and validation of the Reading section, and of these there are not many. The most relevant study not published by the ETS on the TOEFL Reading section is a case study by Peirce (1992, 1994) on the process of development and revision of a set of reading items for operational use. Although the study was not published by the ETS, Peirce was an employee at the test development department at ETS when she conducted the study, and thus she provides a participant perspective into development. A study that is illuminating for the validation of the TOEFL test is Bachman et al.'s investigation into the comparability of TOEFL and the Cambridge First Certificate of English, the *Cambridge-TOEFL comparability study*

(Bachman et al. 1988, Ryan and Bachman 1992, Bachman et al. 1995). I will briefly present these studies below. I will follow the pattern used earlier in this chapter and mention, for each study: the purpose of the study, data or materials, methods, and the main results.

Peirce (1992, 1994) conducted a case study on the preparation of one reading passage and its associated items from start to publishable state. The purpose of her study was "to demystify the TOEFL reading test at a descriptive and theoretical level" (Peirce 1992:666). According to Peirce, this represented a gap in the research published on the TOEFL test up to then, since previous research had "[not] addressed … basic assumptions about what the TOEFL actually tests, why, and how." Peirce's main material consisted of her participant knowledge of the development of TOEFL Reading tests, and a record of successive drafts of one particular reading comprehension passage and related items. Each successive draft of the items was accompanied by a record of the test reviewers' comments and the revisions Peirce had made in response. The near-final test form also included information on the statistical analysis of pretest data. Furthermore, Peirce analysed her own work using applied linguistic theory. Her results showed the technical care with which the TOEFL Reading items are written and raised theoretical questions about the principles followed in test development and test use.

Since there are two published versions of Peirce's case study (1992, 1994), a note is due on my use of them. The author states in the 1994 paper that large sections of it are drawn from the 1992 one. The differences are that Peirce (1994) reports the development of all the items for the reading passage while Peirce (1992) only includes two as illustrations. Peirce (1992), on the other hand, also discusses the theoretical implications of the development practices and the institutional power of TOEFL on what is assessed in the test, which the 1994 chapter does not do. Where overlap exists, I will refer to the earlier paper in the analyses.

The *Cambridge-TOEFL comparability study* (Bachman et al. 1988, Ryan and Bachman 1992, Bachman et al. 1995) was a broad program to find an accountable way of establishing the comparability of two different examination systems. The researchers considered the most important aspect of comparability to be that of the abilities measured, and they employed two complementary approaches to investigate construct validity: qualitative analysis of the content of the two tests and quantitative investigation of patterns of relationships among section scores. They also investigated item statistics where available (Bachman et al. 1995:18). The data that the project used was an "institutional", ie. recently disclosed, TOEFL test together with

a Test of English Writing and a Speaking Proficiency English Assessment Kit, and the Cambridge First Certificate of English (FCE) administered during one operational testing round. For the main trial, performances were gathered on eight sites around the world with a total sample size of approximately 1,450 examinees. The comparability study investigated tests in all the four modalities and concluded that although there were differences in content and format between the tests, there many similarities as well. Detailed results were different for the different test sections. In the discussion in the present chapter, I will only use those parts of the Cambridge-TOEFL comparability study which concerned the reading section.

In the next sections of the present chapter, I will analyse the studies published about the development and validation of the TOEFL Reading section. I will organise the discussion by the categories of activity I identified in Figure 3. These are operational test development, monitoring and maintenance, and ongoing empirical validation. There were no studies published on the operational administration of the TOEFL test as I defined the category in Figure 3, probably because this is an implementation issue rather than a research one. This category is therefore not included in the report below. Analyses of test length, which arguably concern test administration, will be discussed under operational test development because this is the way in which the studies were conducted, not as analyses of existing test administration practices but as possible changes in the standardised administration procedures.

6.3 Operational development of the TOEFL Reading test

There are two perspectives into the operational development of the TOEFL reading test available from the literature. One describes the actual operations involved in the writing and revision of test items, and the other discusses the principles followed in the construction of the test. Furthermore, the reports that discuss the revision in test form in 1995, where the vocabulary items were incorporated in the reading passages, constitute a conceptual group of their own. This is because they describe a change in the test rather than standard test development.

6.3.1 Item writing and revision

The procedures followed in the writing and revision of items for TOEFL Reading test are described in a case study by Peirce (1992, 1994). Peirce (1992:669) explains that the writing begins with item writers external to ETS,

who find appropriate passages and develop 6-7 items for each of them according to detailed specifications. The drafts are sent to ETS, where a test writer employed at the test development department receives them and, in a number of stages, converts each passage-and-items combination into "a publishable pretest set" (p. 669). This means that the ETS test writer reads the text, works on the items, and then submits them into a comprehensive review process. The test writer works on the items rather than the text itself because, in the interests of authenticity, ETS discourages editorial changes to the text (Peirce 1992:675).

Once the test writer is satisfied with the testlet, the review process begins. This consists of two main parts: a series of test reviews by "approximately six different test development specialists" at ETS (Peirce 1992:672) and a pretesting process where the pretest set is inserted into a final TOEFL form to gather scoring data (p. 667). An item analysis is conducted to determine the difficulty and discrimination of the pretest set and to identify any items which do not work properly. Such items are then revised or discarded and the statistical information is used when the set is incorporated into an operational TOEFL test (Peirce 1992:667).

The test review process prior to pre-testing was developed at ETS to help avoid potential problems with the items which might ensue if only one person prepared the items (Peirce 1992:672). The coordinating responsibility for developing a set remains with one individual, however, in that each reviewer's comments are returned to the test writer, who makes some revisions but who can also defend the original solution and not make a recommended revision. In her case study, Peirce refers several times to the statistical analysis following pretesting, which can help resolve questions on which the test writer and reviewers have disagreed. All the reviews are kept in a folder at ETS until the test is ready for publication, so that the development of any item can be checked by any reviewer if they wish (p. 673). These files enabled Peirce to conduct the case study.

The review process begins with a Test Specialist Review (TSR), where another member of the TOEFL test development team takes the test and makes comments. After revisions, the draft is reviewed by the TOEFL coordinator, then by two stylistic editors, and finally by a sensitivity reviewer, who checks the passage and the items for potentially offensive material (p. 673). Next, galleys are made of the set, and once these are checked at the Test Development department, the set is published for piloting in an operational test form. The pretest data is analysed at the Statistical Analysis department and the results are forwarded to the Test Development department. The test developer then decides if any items need

to be revised or discarded before the set can be included in a final form (p. 667).

Peirce explains that to judge how well an item has worked in a pretest, the standard practice at ETS is to compare the way that each item worked with the candidates' performance on the total for Section 3 (1992:678). The sample that has taken the pretest is divided into quintiles based on the total score for Section 3 and the number of candidates who chose each alternative is recorded in a table. Biserial correlations are calculated for each item. As the pretest items do not influence the candidates' total scores, there is no problem of the item being included in the total score.

The criterion that ETS uses for accepting an item is a biserial correlation of .5 (Peirce 1992:678). Values around or below this critical value indicate that the item does not work in the same way as the rest of the test. However, the application of this criterion is not mechanical, as Peirce illustrates (1992:680). The biserial correlation for one of her items was .55, but when she examined the distribution of the correct and incorrect responses, she found that too many candidates in the top two quintiles chose one of the distractors. One of her reviewers prior to pretesting had indicated that this might be a problem, and the pretest statistics now validated this comment. Peirce concluded that this distractor would have had to be modified if the test had been sent forward for operational use.

Peirce's (1992) study includes an account of the principles that she followed when developing tests and items. Her three main principles were to use the candidate's time efficiently, to help the candidates orient themselves to the text, and to make sure the items were defensible (pp. 670-672).

Using the candidates' time efficiently meant, for Peirce (1992:670), that she should develop as many items as the passage could sustain and delete any additional portions of text if this did not disturb coherence. Moreover, she should use closed rather than open stems, that is, stems that end in a question mark, so that the candidate would not have to re-read the stem with every alternative. An observation of the set that Peirce used in the case study (1992:690-691) indicates that she did not follow this latter principle too rigorously, as six of her nine items have open stems.

Peirce's principle of helping the candidates orient themselves to the text arose out of the TOEFL practice that reading passages do not have titles or contextualizations (1992:671). She thus attempted to create a first question which focused on the main idea. Other strategies for helping the candidates orient themselves to the text included the aim to present items in the order of the text and the use of line references in questions as much as possible.

Making sure that the items are defensible in combination meant doing justice to the content and level of difficulty of the text. To Peirce, this represents the art of test development, since importance and complexity are individual judgmental decisions. If the text is difficult, the items should be so too, so that the candidates have the opportunity to demonstrate their advanced understanding. At the same time, the items should be independent and each of them should focus on a different idea in the main text.

Peirce's list for item defensibility resembles standard checklists for technical adequacy of multiple choice items (p. 672). The stem and key should be unambiguous, options should not overlap logically, distractors should be plausible but not potentially correct, the key should only be identifiable if the text is understood and not without reading or understanding it, and all the options should be structurally and stylistically similar, so that none could be eliminated on this basis rather than with reference to the text.

In addition to item writing, each of the TOEFL test writers also acts as a reviewer for other test developers' items (Peirce 1992:672). Reviewing is also guided by principles, and in her article, Peirce (1992:673) also characterises her own style of reviewing. She says she was particularly concerned about items which were potentially ambiguous, ie. had more than one potential key or perhaps no clear key at all. She felt less strongly about stylistic weaknesses or implausible distractors. The reviewer comments made by her colleagues at ETS that Peirce reports in the case study (1994:47-54) reveal that all ETS test developers seem to pay attention to similar matters, though perhaps in different proportions. Peirce received comments on matters such as stylistic differences and typography but also on logical overlap between options and potential for extra difficulty because of the language in the item being more difficult than the language in the passage.

Summing up her case study, Peirce states that her account shows how the two kinds of feedback, comments from colleagues and statistical information from pretesting, combine to constitute "a complex set of checks and balances" with which the testing board ensures that the test is technically adequate (1992:680-681). Moreover, worth noting is the commendable practice of keeping careful record of all comments and revisions to an evolving testlet as a standard test development procedure.

However, Peirce (1992:681-684) herself raises some questions about the principles that she followed especially with regard to authenticity and test validity. She argues against the authenticity principle of not changing the text. Once the passage has been removed from its original context, faithful

replication of the original wordings does not guarantee authenticity because the original textual context is missing. Furthermore, the social meaning of the text when it is presented as a reading passage in a TOEFL test is governed by the testing context. If the examinees read the same passage in some other context, they would also read it for other purposes. Peirce (1992:683) argues that the test context predisposes the reader-as-examinee to read the text and the items for the meaning that the test writer has intended regardless of the possible textual meanings that might be available from the source text. Thus, the meaning of the text in a TOEFL test is different from the meaning of the same text elsewhere whether the exact wordings of the original text are used or not.

As regards validity, Peirce (1992:684) points out that the criterion that she used when she judged the acceptability of an item, a point biserial correlation of at least .50, was self-referential. It meant that the item fitted in with other TOEFL Reading items, but this did not guarantee that it adequately represented the reading construct. The criterion ensured adequate measurement properties for the instrument, but construct-oriented studies would be needed to show what the instrument measures.

6.3.2 Test construction

The procedures followed at ETS to construct entire operational test forms and the Reading test within them are discussed by Chyn, Tang and Way (1995). The study is focused on a possible change in the test construction procedures, and it is written from the perspective of the examination board's policy initiative.

Before Chyn et al.'s (1995:1) study, TOEFL test forms had been constructed at the Test Development department by employees who used a combination of statistical and content criteria and human judgement to guide their work. The statistical criteria had been based on classical test theory. Chyn et al. (1995) studied the possibility to use an Automated Item Selection Procedure (AIS) to help construct final forms of the TOEFL test. This would enable test construction on the basis of Item Response Theory (IRT) data rather than classical item statistics. At the same time, it would enable the use of the IRT-based test information function to evaluate the measurement quality of the whole test (Chyn et al. 1995:1). Previous studies on simulated data had shown that test construction efficiency in terms of time and cost had increased with the use of AIS and that the resulting parallel tests had shown greater content and statistical consistency. The purpose of Chyn et al.'s study was to investigate whether this would be true with real TOEFL test forms.

Chyn et al. (1995:5) developed an automated test construction procedure that combined statistical and content rules related to test development. They selected an appropriate IRT-based information function for the test and developed IRT-based statistical specifications. Together with the test development department, they developed a set of content rules for test construction. The rules concerned item properties such as format, difficulty, skill focus, topic, type of function or structure tested, gender appropriacy and key distribution (Chyn et al. 1995:26). After an iterative process of rule creation and tryout, the final model included 120 rules for TOEFL Section 1, 87 rules for Section 2, and 49 rules for Section 3. The rules were weighted to indicate that some were more important than others.

Two TOEFL test forms were created using the AIS and the forms were submitted to a test review process that parallelled that of the development of a set of items discussed in the previous section. That is, the test assembler reviewed the forms proposed by the AIS and revised or replaced individual items that were not appropriate (Chyn et al. 1995:14-15, 26-27). This was followed by a test specialist review, a test co-ordinator review, and a mechanical layout review. The number of changes made at each stage was recorded.

The results indicated that Section 3 had the greatest number of revisions and replacements, but this was partly due to the larger number of items in this section. The time spent on the AIS review process was compared with an average of test developers' assessments of how long it took to assemble a TOEFL test form manually. It was found that the use of the AIS made the test construction quicker for Sections 1 and 2 but only potentially so for Section 3. The evidence was inconclusive because the time spent on reviewing and revising one of the two Section 3 forms was much shorter than the traditional method while the time required for the other form was much longer. For all the sections, the degree of statistical parallellism between the forms assembled with the help of the AIS was better than that of manually assembled tests. The results also showed that the TOEFL item pools allowed the use of IRT-based statistical specifications in test assembly even if it had been suspected that this more complex criterion would be impractical as compared with the equated item deltas from classical test theory that had been used previously (Chyn et al. 1995:31).

As a part of the study, Chyn et al. (1995:28-29) surveyed staff reactions to the AIS assembly of TOEFL tests. The reactions were generally positive although some drawbacks were also detected. The positive features noted by the test developers were time efficiency, help with

the balancing of routine content characteristics of tests such as key position and gender appropriacy, increased objectivity when analysing a test, and encouragement of regular review of the contents of item pools. The drawbacks included a lower degree of ownership for the test form felt by the test assembler, lower cognitive demand of the assembly task and, for some respondents, the neutral quality of the AIS-based test forms instead of the more balanced content that a form constructed by traditional means would entail.

A significant point raised for discussion in the Chyn et al. (1995) study was the quality of the item pools that are used in test assembly. The number of changes needed after the AIS had been applied was too high, and a major implication of the study for test development was that quality control should be improved (Chyn et al. 1995:32). The monitoring of the content of TOEFL item pools should also be changed to fit the AIS requirements. As a side product of this discussion, the concerns raised about item selection and replacement show that TOEFL items are coded for a rich range of content and statistical properties, all of which can be potentially used when items for test forms are selected with complex automated algorithms. As revisions and replacements to the initial AIS form were made, it appeared that there were some abstract or vaguely formulated content considerations in the test content specifications which were not reflected in the AIS rules but which human judges were nevertheless able to apply when they selected items for a test (Chyn et al. 1995:29). Such rules would have to be worded, coded as item properties, and included in the AIS content rules if the decision was made to use AIS in future test construction.

The discussion above shows that both statistical and content concerns are attended to when final TOEFL forms are constructed. With the introduction of automated item selection, it is possible that routine application of content criteria becomes more formalised and potentially more conscious. The statistical successes and content challenges that Chyn et al. (1995) experienced when they implemented the AIS showed that the test developers had better control over the measurement properties of the test than the content specifications. However, the work on the content categories showed that the researchers considered it an important concern. The connection between the psychometric definition of the construct and the content properties of the items was not made in the study. The development of a “content information function” similar or analogous to the IRT-based test information function that was used as a basis for evaluating

the performance of the AIS may be a long way off, but the content categories may offer a way of developing it.

6.3.3 Development of the all-passage TOEFL Reading section

In June 1995, the format of TOEFL Section 3 was changed. The vocabulary items, which had formerly been discrete items in the context of a sentence, were incorporated into the reading passages. Schedl, Thomas and Way (1995:1-3) report on the steps involved in the change. Henning (1991:14) provided some supporting information, namely that the psychometric quality of passage-embedded vocabulary items was at least as good as that of discrete vocabulary items, but he did not discuss the impending change in the test. Thus, the test development activities related to the all-passage reading section are only reported from the Test Development group's point of view in Schedl et al. (1995).

Schedl et al. (1995) begin their report with a brief discussion of the considerations that led to the change. They classify the format that was traditionally used in TOEFL vocabulary items as discrete point testing. They state that this approach to testing was "based on an understanding of language proficiency as a set of linguistic abilities which could be separately measured (phonological, syntactical, lexical)" and that this understanding was common at the time when the TOEFL test was developed in 1963 (Schedl et al. 1995:1). With theoretical developments in language learning, language testing, and reading theory, especially the introduction of communicative competence, these items fell into disfavour because "communicative tests do not focus on measuring discrete aspects of language performance since, in authentic language use, grammatical, phonological and lexical knowledge do not manifest themselves independently". Rather, knowledge of vocabulary is required in longer contexts which give more clues about the meaning of the words and phrases than a single sentence. Schedl et al. (1995:1-2) explain that research information about the effects of context is conflicting, but that context-embedded items have better face validity and are assumed to have better washback effects. The assumptions are nevertheless presented as possible motivations for the change in the test.

The Test Development group conducted a pretest study in 1989 with two alternative formats of a vocabulary test, one where the items were tested in the context of a single sentence, and the other where the items were embedded in a passage. The passage-based items were found to be more acceptable in terms of face validity, rich availability of context clues, involvement of reading as well as vocabulary, and likelihood of beneficial

washback (Schedl et al. 1995:2-3). The study showed no great psychometric disadvantages with the passage-embedded vocabulary items either. Following this, in 1990, the Committee of Examiners began to explore ways to incorporate this change in the operational TOEFL. A larger-scale trial was arranged at a number of English Language Institutes with an experimental 54-item all-passage reading test and an "institutional" TOEFL test, ie., a TOEFL test form that has been recently disclosed and that the institutes had bought for their own use. The institutional test followed the old format with separate sections for reading and vocabulary. The statistical analyses "indicated that the new format was reliable and that items fell within the current range of difficulty for TOEFL" (Schedl et al. 1995:3). However, the new test appeared to be significantly speeded. This motivated further study.

Schedl et al. (1995) investigated the speededness of the proposed new Section 3. They administered three versions of the new test under three timing conditions. All test versions included the same six passages, with either 48, 54, or 60 items on them. The testing times were 50, 55, and 60 minutes. They found that the number of items included in the test was not as influential as the number of passages, and that to safely count the test as non-speeded, the inclusion of six passages would require a testing time of 60 minutes or more. They recommended the inclusion of five passages in the revised test with a minimum of 55 minutes time allowance (pp. 15-16). They conducted equation analyses between the old and the new reading section using the scores from five rather than six passages in the new test form and concluded that the equation was possible and the five passage test had adequate reliability (p. 14).

Furthermore, Schedl et al. (1995) conducted a preliminary analysis of dimensionality in the old Vocabulary and Reading section to test if the removal of a separate vocabulary section would change the psychometric characteristics of the test. They found that discrete vocabulary items did not form a separate dimension, but that there were traces of possible speededness effects in the data (1995:28). They recommended further studies into appropriate test length and time limits for the revised Reading section (1995:31). I will discuss the studies conducted so far under the section for test monitoring and maintenance.

The reason for the change in the format of the Reading section reported in the Schedl et al. (1995:1) study was that the examination board wanted a new test that reflected more desirable construct properties than the earlier one. The qualities of the new test investigated in Schedl et al. (1995) and in the previous trials that they summarised were related to the

measurement properties of the test and to the threat of speededness. The researchers did not mention investigations into whether the construct description for the test changed nor did they indicate this as a concern. On the one hand, this indicates a commitment on the part of the examination board to the psychometric construct that the test embodies: the measurement properties of the test were shown to be equal to those of the earlier version, so the question about what was measured in the test did not arise with any more urgency than it had before the change. On the other hand, this begs questions about what precisely the change meant and why it was made if the nature of the skills assessed in the test is not a concern for the test developers. Spolsky's (1995) answer would be that the change was made because the audience demanded it. The present study cannot provide answers to this question.

6.4 Test monitoring and maintenance

Studies on test monitoring and maintenance include publications that discuss routine procedures by which the test developers examine the quality of their tests and assessment procedures and monitor needs for revision. The implications of the study on test construction discussed above (Chyn et al. 1995) arguably belong to test monitoring as well, but since they were discussed above, I will not repeat the discussion here. I will, however, report on the studies that have been published on the regular monitoring of the quality of the TOEFL Reading section. These concern pretesting and equation procedures and test reliability.

Way, Carey and Golub-Smith (1992) investigated the pretesting and equation procedures of the TOEFL test. Some time before their study, a new quality check on the equating procedures had been implemented. This involved routine checking of observed item parameters against those estimated at the pretest stage. Way et al. (1992) used data available from this procedure to explore how the differences that are detected between pretesting and operational use might be explained. Regarding the reading section, Way et al. (1992:8) found that an important factor seemed to be the position of the reading passage within the section. Items which had been pretested near the end of a test form were likely to have different item characteristics if they were used towards the beginning of the reading section in an operational form, and vice versa. The authors suggested, referring to other studies in the ETS Research Reports series (Bejar 1985, Secolsky 1989), that this may be related to the possible speededness of TOEFL Section 3. They recommended that the instructions for compiling

final forms for the Reading test should be slightly modified to keep the relative position of each operational passage as close as possible to the position where the passage was pretested.

In their analysis of the reliability of the TOEFL test, Wainer and Lukhele (1997) paid specific attention to the reading section where items are bundled in testlets. Their overall results repeated previous findings that the TOEFL is an extremely reliable test even if item dependence within a testlet is assumed. However, they pointed out that this result concerns the *total score*, not the subscores for listening or especially the new, all-passage reading section. They showed that once item dependence is taken into account, which had not been done in previous studies, the reliability coefficient for the reading test was reduced. To reach the levels of reliability that the test developers thought they had on the basis of the previous studies, the reading test would actually have to include seven ten-item passages of the new kind. This is a test length that the TOEFL cannot deal with practically, so the board has to accept lower reliability coefficients for the reading comprehension sub-score.

Wainer and Lukhele (1997:10-11) pointed out that the reliability coefficient of the new reading section when item dependence is taken into account, .86, is lower than earlier but still perfectly acceptable. It was fine to "trade off a little reliability in order to obtain a test structure that characterizes the domain of the test more accurately," and this is what the testing board gained with the change. Wainer and Lukhele's point was simply that the test developers should recognize the reduction in reliability, especially if subscores will be used more broadly, for instance for diagnostic purposes.

The studies published in the TOEFL Technical Report series that concern methods of test equating and test construction are all relevant to, and motivated by, test monitoring and maintenance. The reason why Way and Reese (1991:1) conducted their study of the use of one-parameter and two-parameter IRT estimation models for scaling and equating the TOEFL was that if it was possible to use the simpler models, smaller sample sizes would be required in pretesting and the analysis of the results would be quicker and cheaper. However, the model-data fit and the comparability of the measurement properties of different forms were best by all statistical indicators when the three-parameter logistic model was used, which meant that the less demanding statistics could not be adopted. Tang et al.'s (1993) analysis of programs used in IRT-based scaling and equating was similarly motivated by a wish to find the best and most efficient means for conducting analyses. Such studies on test monitoring are primarily

motivated by economy and efficiency and are presumably conducted on areas that are considered important and/or expensive by examination developers. Assuming this is correct, the areas that seem to be expensive and important to the TOEFL test developers are efficiency of test development procedures, the test's psychometric properties especially in terms of reliability and test equation, and the test's focus on appropriate skills as evidenced in Chyn et al.'s attention to test content. The aims of these studies follow Spolsky's (1995:318) analysis that at least some of the TOEFL research agenda is "product-, market-, and profit-oriented."

6.5 Empirical validation of the TOEFL Reading test

Most of the studies discussed above are related to the validity of the TOEFL Reading section in one way or another. The research discussed below is centrally concerned with what the TOEFL Reading section measures. As discussed in section 6.2, many of the studies that are categorised in *The Researcher* under construct validity approach the topic from a perspective where the TOEFL total and section scores are givens and where the main construct concern is whether reading is one of the identifiable secondary dimensions reflected in the score data. Studies that investigate the construct in more detail also exist, however, and these constitute the bulk of the discussion below.

Duran, Canale, Penfield, Stansfield and Liskin-Gasparro (1985) described the content characteristics of one TOEFL test form in terms of an exploratory framework of communicative competence. They developed the framework with reference to contemporary models of communicative competence all carefully referenced (Duran et al. 1985:6-11) and applied it with the intention of indicating both what the test measured and what aspects of proficiency it did not measure (Duran et al. 1985:1). They developed a communicative skills checklist and analysed each of the TOEFL sections to see which skills it covered; although they noted (1985:12-13) that a checklist is only "an operationally oriented list of discrete skills entering into communicative competence and the communicative process" rather than an accurate depiction of the integrative nature of the construct, it nevertheless helped analyse the skills that were intentionally tested in the TOEFL test. Throughout the report, the authors emphasized that their analysis was preliminary and exploratory and that confirmation of its findings and refinement of its instruments would be needed in the future. All the test items were coded independently by two of the researchers after some initial consultation, but the codings of one

researcher were reported on grounds of greater detail and expertise (Duran et al. 1985:20). The researchers also analysed the requirements of the TOEFL test in terms of Bachman and Palmer's (manuscript) test performance factors, which were: psychophysiological skills in test taking, representation of knowledge, language use situations, context and message, artificial restrictions, monitoring factors, affective factors, and strategic factors (Duran et al. 1985:21). A preliminary discussion of the relevance of the properties to the TOEFL test was presented but the properties of the test were discussed in generic terms and not with reference to different test sections. The test was also evaluated in terms of its relevance to academic and social language use contexts. Ratings on a scale from 1 to 3 were given on three criteria: relevance of content and language to everyday college life, relevance of topics to formal instruction at college level, and as concerns the listening section, to social naturalness (Duran et al. 1985: 21-22). Two raters rated all TOEFL items on these dimensions independently and the scores were averaged over each test section and each item type. In a final analysis, two researchers evaluated the difficulty level of the TOEFL items on the Interagency Language Roundtable (ILR) scale. The researchers "individually reviewed the [TOEFL] form and then jointly discussed it" (Duran et al. 1985:23). The findings reported in the study represent a joint summary of the observations made. The instrument categories are reported carefully and the test form analysed is reproduced in the appendix to the study, so that the analysis can be replicated.

The main findings of the analysis of the Reading section in terms of the communicative checklist were that the reading passages offered the examinees a rich sample of language and that this enhanced the communicative nature of the test (Duran et al. 1985:38). The variation in sentence structure was commended, whereas variation in patterns in the rhetorical and semantic organization of ideas as in classification or cause and effect seemed to be largely missing (Duran et al. 1985:38-39). The authors suggested that this may reflect a deficiency in the categorisation if it does not reflect textual structures typical of academic texts or a problem with passage length in the test such that different textual structures are not reflected in them. They judged the topics of the reading section fairly typical of academic classroom content, but they also noted that the "open stem" format of many of the test items was artificial. That is, the examinees were unlikely to meet the type of "complete this sentence" tasks in real life (Duran et al. 1985:39).

In the concluding statement to the communicative checklist, Duran et al. (1985:40) contended that the findings they reported were only based on

one form of the test. The method would be more useful, they suggested, if it were applied to a range of test forms, after which assessments could be made of the stability of components and skills required in different forms of the TOEFL test.

The main finding of the ratings of content relevance to academic language use was that the Vocabulary and reading section was not particularly specific to academic contexts. The ratings averaged 2 on the 3-point scale. The definition for a judgement of 2 was that the materials "might be relevant to academic life or to college level academic content materials, but that there was no clear and compelling evidence to assert overwhelmingly that they were on the average" (Duran et al. 1985:47). The authors suggested that this was because of "the absence of information concerning the pragmatic meaning that could be attached to the content meaning of items." This statement is very similar to Peirce's (1992) assessment that the reading texts lose their original contextual meaning when they are placed in a TOEFL Reading test.

Duran et al.'s (1985:55-56) evaluation of the difficulty of the Reading section on the ILR scale was that the passages varied in difficulty from level 3 to level 4 but that the questions focused on level 3. Their judgement was based on ILR level descriptors, particularly those on level 4 of requiring the ability to follow unpredictable turns of thought and to recognize professionally relevant vocabulary that can be assumed to be familiar to educated nonprofessional native speakers. Their analysis indicated that such skills were not demanded by the reading items. They concluded that this may be highly appropriate since the test was intended for non-native speakers and items at level 4 might present difficulties for some native speakers (Duran et al. 1985:56).

Duran et al.'s (1985:60-62) conclusion from the content analysis of the TOEFL test was that it was valuable as an instrument for assessing non-native speakers' language proficiency and that its reading section provided rich stimuli for communicative language use. They recommended that in addition to reliability, validity and practicality, the test developers should consider the acceptability of the test and its feedback potential when they make test development decisions. They defined acceptability as "the extent to which a test task is accepted as fair, important, and interesting by both examinees and test users" and feedback potential as "the extent to which a test task rewards both examinees and test users with clear, rich, relevant, and generalizable information." They recognized that such considerations are often given lower priority in the construction of high stakes tests but they pointed out that in the interests of serving the examinees and educators

who prepare them, perhaps they should not be ignored altogether (Duran et al. 1985:62). They also proposed a research agenda for TOEFL program activities that included continued study of the content characteristics of TOEFL items, development of research on thematic presentation of items and new item formats while considering the possible drawbacks in psychometric properties and examinees' test performance, new approaches to assess speaking and writing directly, incorporation of technology in terms of innovative measurement and adaptive testing, and broadened validity research. Under this title, Duran et al. (1985:66) particularly emphasize the importance of studying empirically the construct assessed in the TOEFL test through the identification of appropriate criterion tasks and performances, development of performance measures on these tasks, and comparison of TOEFL against such criteria. In hindsight, it is easy to see that this program heralded the development of the TOEFL 2000 project.

At the time when the Duran et al. (1985) paper was nearing completion, a conference was arranged at ETS to discuss the TOEFL program in relation to the notion of communicative competence. Theoretical papers were prepared and circulated in advance, and the discussion included both prepared reactions and general discussion. Many of the issues raised concerned general approaches and the testing of writing and speaking. I will only refer to those parts of the presentations that specifically concerned the Reading items. Bachman (1986:81-83) concurred with the Duran et al. (1985) conclusion that the reading section of the test offered the greatest potential for the assessment of communicative competence. He considered the test situationally authentic in that academic reading most generally entails interaction between an individual reader and an academic text. He raised the question of how necessary it was to observe the psychometric requirement of local item independence when theoretical treatments of the nature of communicative language ability implied that performance on items was, and should be, integrated (Bachman 1986:84). He also questioned the reliance on four-option multiple choice only and proposed that the introduction of more creative test procedures and formats might be seen as a challenge to psychometricians to provide new models that better fitted the nature of the abilities that language testers desired to assess (Bachman 1986:85). Oller (1986:145-146), approaching the evaluation of the TOEFL reading section from a rich view of communicative competence as a creative resource, criticized the topics of the reading section as dry and academic. He argued that elements of disequilibrium, doubt, puzzlement, surprise, or conflict would motivate the text and the reader's reading of it, (Oller 1986:145). This criterion of text selection is

related to Oller's notions of authenticity of texts and meaningfulness of questions, which build on the reader's interest and attention. He recommended controlled studies of psychometric properties of tests with varying degrees of coherence and authenticity (Oller 1986:149) and echoes Bachman's call for investigations of a varied range of item types. Both contributions illustrate the range of test development considerations that are opened when the basis for asking questions is broadened from psychometric measurement properties. Nevertheless, both authors also stress the importance of investigating the effects of any new conceptual developments on the psychometric properties of the test.

Freedle and Kostin's (1993a, 1993b) study of predicting the difficulty of TOEFL reading items and Boldt and Freedle's (1995) re-analysis of the data were discussed in Chapter 4. Since this study was part of the empirical validation of the TOEFL reading section, a brief summary of its content and implications is due here. The study involved a textual analysis of reading passages and items in order to discover a set of features which would explain the difficulty of the items. The researchers were concerned about criticisms that multiple choice tests of reading might not test reading but ability to deal with the questions, and they were able to show that the textual features of the passages and passage-item overlap were clearly more significant in predicting item difficulty than the textual features of the items. The proportion of variation explained was 62% at best, however, and Boldt and Freedle's (1995) re-analysis suggested that it may have been inflated. The conclusions of the latter study did not challenge the original study's construct argument that features of the reading passages were better predictors of item difficulty than features of the items, but a somewhat disconcerting finding was that the variables that proved useful in Boldt and Freedle's (1995) analysis were largely different from Freedle and Kostin's (1993). The report on the re-analysis regrettably contains no construct-related analysis of what proportion of these variables were related to passages, items, and passage-item overlap. Boldt and Freedle's (1995) study appeared in the Technical Report series, and its motivation may have been technical or economic rather than construct-related. It is nevertheless coded as a report that concerns the construct validity of the Reading section (ETS 1999:2). The grounds must be different than the concepts used to explain what the test tests.

Schedl, Gordon, Carey and Tang (1996) investigated whether reading items designated as testing "reasoning" formed a separate measurement dimension in the TOEFL reading section. Their review of previous studies of the possible existence of reading subskills indicated that the evidence was

inconclusive and that it was certainly possible that the reasoning items might comprise a unique trait within reading. The possible subgroup consisted of items testing "(1) analogy, (2) extrapolation, (3) organization and logic, and (4) author's purpose/attitude" (Schedl et al. 1996:3). If these items constituted a subgroup, this might imply that the specifications for the reading test would have to be changed so that a set proportion of all TOEFL reading tests would measure this dimension. If they did not, the test developers would be justified in continuing their practice whereby these item types are considered to contribute variety to the *overall* assessment of reading (Schedl et al. 1996:3).

With scores from more than 1000 candidates per test form, Schedl et al. (1996) applied a test of essential unidimensionality for each of the 10 test forms involved in the design and analysed the data with nonlinear factor analysis. They found that the reasoning items did not comprise a second measurement dimension (Schedl et al. 1996:9). They explained, however, that this does not mean that conceptually distinct subskills do not exist, all that the result indicates is that a separate latent ability trait is not needed to characterise the performance differences between these candidates. From a test development point of view, the implication of this study thus was that rules of test development did not need to be changed. From a validity point of view, it indicated that the measurement construct was unidimensional. Means for investigating the nature of the construct assessed in other than numerical ways were not discussed.

Regardless of the finding that reasoning items did not form a distinct measurement dimension, Schedl et al. (1996:10) found a minor secondary factor in their data. Significant *T* statistics for exploratory one-factor analyses indicated that essential unidimensionality was rejected for all the tests investigated. Exploratory two-factor analyses indicated that, for the Reading section, all the items that loaded on the second factor more clearly than on the first were associated with the last two text passages in the test. The authors interpreted this to mean that the factor was related either to passage content or to passage position, not to passage difficulty. The study thus indicated similar tendencies as Way, Carey and Golub-Smith's (1992) investigation into parameter variance between pretests and operational tests if item position changes. Schedl et al. (1996:10) noted that since the second dimension was always related to the last two passages of a reading test form, the dimension might be explained by time pressure or examinee fatigue. They suggested that shorter versions of the reading test where pretest items are not included could be analysed to see if a similar end-of-test effect could be found there. Regarding passage content, the researchers

proposed that the passages in their design could be analysed for content differences and a new design could be made with these and other passages on similar topics to investigate whether passage content constituted a separate measurement dimension. The results of such studies might inform future test design. In terms of validity, studies of speededness can be categorised as ones investigating rival hypotheses for score explanation and studies of content effects as investigations into the nature of the measurement construct.

Boldt, Larsen-Freeman, Reed and Courtney (1992:1) reacted to a proposal from score users that TOEFL scores could be more meaningful if it was possible to describe verbally what the score levels mean. They investigated the possibility by comparing the listening, reading and writing sections of the TOEFL with the American Council on the Teaching of Foreign Languages' (ACTFL) rating system (ACTFL 1986), which uses such verbal descriptors for different score levels. Some 84 teachers of English as a second language from seven different colleges in the eastern United States were asked to rate as many of their students as they could on the ACTFL scales for listening, reading and writing. The number of students rated was 369 for reading and 405 for listening and writing altogether, while the group sizes within institutions varied between 29 and 102. The teachers were not trained for the rating work, but as Boldt et al. (1992:4) note, there are no certification procedures for ACTFL raters in the three skills studied but only speaking, which was not investigated. The students also took an institutional TOEFL test and the researchers compared the distributions of the ACTFL ratings across ranges of TOEFL section scores.

Boldt et al. (1992) faced interesting challenges in operationalization because the ACTFL scale is not directly oriented towards numerical expression but uses verbal descriptors such as "Novice-Low", "Novice-Mid" and "Novice-High" and because the logic of student assessment through the ACTFL scales relied on teachers' knowledge of the students. This meant that it was not possible to design a double rating format that would have made the study of rater reliability or rater severity straightforward. Boldt et al. (1992:7) solved the scale issue by using two different numerical expressions for it, an equal interval scale from one to ten and an unequal variant developed in an earlier study by Lange and Lowe (1987). They adjusted for rater severity through what they termed a "football correction" familiar from sports betting contexts; the procedure is not reliable but it is the best possible one in a situation where direct pair comparison data does not exist for all pairs (Boldt et al. 1992:4-6). When

more than one rating per student existed, Boldt et al. (1992:9-10) used these to assess rating reliability. They found reliabilities of around .60 for the whole sample, but they also found variation between different institutions (Boldt et al. 1992:10-11, 22). The reliabilities were used in assessing the significance of adjusted correlations between the ACTFL ratings and the TOEFL scores. For the Reading section, for example, although the corrected correlation was .77, its maximum was in the low .90s, which indicated that not only was reliability a problem but the ratings and the TOEFL section proved not to be entirely parallel measures of proficiency (Boldt et al. 1992:10).

The result of the Boldt et al. (1992) study in relation to its original motivation was a set of tables that indicated the range of ACTFL ratings which students at different TOEFL score ranges were assigned. The spread of the ACTFL ratings for any score level was quite broad, for instance those who scored below 40 on the 60-point TOEFL Reading scale were assigned ratings from Novice-Mid through the whole Intermediate range to Advanced, and those who gained a TOEFL Reading score between 45 and 49 were rated from Novice High through Intermediate to Advanced Plus on the ACTFL (Boldt et al. 1992:45). It was thus not possible to adopt the ACTFL scale descriptors for explaining TOEFL score ranges, but Boldt et al. (1992:13) proposed that the information about the spread of levels was still useful for admissions officers. The study raised questions about the numerical operationalization of scale-based assessments but the authors were nevertheless able to conclude that in broad terms, the TOEFL and ACTFL tapped similar if not exactly the same skills (Boldt et al. 1992:12). In a concurrent validation sense, they thus considered the reasonable correlations between the two measures to support the construct validity of both. However, they raised questions about intercorrelations of ratings and scores between skills, which meant that the data did not offer strong numerical evidence that listening, reading and writing were indeed distinct as skills (Boldt et al. 1992:13). They propose that if individuals could be found whose scores in different skills truly were different on one measure, they should be tested using other measures so as to see if the skill difference re-appears. The theoretical orientation in this statement assumes a skill construct that should preferably be individual-specific and independent of the context in which it is expressed – in terms of Chapelle's (1998) figure (see Chapter 4 Figure 1), a trait-oriented, person-emphasized conception of ability. In terms of verbal and numerical score definitions, the approach of the study was numerical: although it started from a wish to describe score meaning, the expression it found was a set of score ranges. Descriptions of

the abilities expressed in the scores through the contexts in which performances were elicited was not attempted. Both kinds of studies are needed, but the call for the description of a contextualised construct was not raised in the discussion of the Boldt et al. (1992) study.

Outside the TOEFL research reports, Bachman, Kunnan, Vanniarajan, and Lynch (1988) analysed the content of one TOEFL test form in a study somewhat reminiscent of the Duran et al. (1985) analysis discussed above. Bachman et al. (1988, 1995) analysed the content of one institutional form of the TOEFL Vocabulary and Reading section as part of the Cambridge-TOEFL comparability study. They used Bachman's (1990) frameworks of Communicative Language Ability (CLA) and Test Method Facets (TMF) to guide their investigation. A fairly rigorous application of the techniques, published in the 1988 article, involved a close textual analysis of the reading passages and items in much the same textual detail as in the Freedle and Kostin (1993) study discussed above.

Bachman et al.'s analysis of the language in the reading section revealed that the TOEFL passages were all academic and that the passages had a linear or sequential progression, reflected among other things in a frequent use of clausal connectives. The illocutionary range of the TOEFL passages in the test form that they analysed was relatively narrow, with just over half of all acts concentrating on "giving information". The researchers' assessment of the kinds of strategic competence tested in the TOEFL reading section was that selection, analysis, and synthesis were represented equally well, while evaluation/judgement was only required on two occasions. Bachman et al. (1988:154) note in their discussion that since their sample for the tests was minimal, one form only, any statements that they make concerning the tests could only be tentative.

For the main publication about the Cambridge-TOEFL comparability study (Bachman, Davidson, Ryan, and Choi 1995), the strategy for the analysis of test content was changed so that the tests were not analysed textually but by expert evaluation. The experts judged how far the different facets of Bachman's CLA were required by the two tests and what the test methods in the two tests were like. The Cambridge test investigated in the main study was the First Certificate. The main result was that the Cambridge and TOEFL tests were more alike than different in terms of abilities measured and test methods used. The only facet of communicative language ability that the experts judged which was significantly involved in all the tests and subtests was vocabulary, the next strongest involvement was perceived for knowledge of syntax and cohesion (Bachman et al. 1995:123-124). The results for the analyses of the reading sections mirrored the overall results,

although the TOEFL reading items were perceived to require a higher level of lexical knowledge than the FCE. The conclusion, as far as any can be made on the basis of one test form, was that the TOEFL and the FCE, their reading sections included, seem to measure language ability rather narrowly, in terms of lexical and grammatical knowledge (Bachman et al. 1995:124).

Ryan and Bachman (1992) investigated differential item functioning (DIF) in the TOEFL and the FCE. They found that the TOEFL reading section included several items which functioned differentially for examinees from an Indo-European and a non-Indo-European background, but they explained that this difference was expected because the first language influences second language knowledge according to a number of earlier studies, which they quote and summarise (Ryan and Bachman 1992:23). The authors attempted to explain the DIF through content variables used in the main Cambridge-TOEFL comparability study summarised above, but they found no clear patterns of task characteristics which might have caused it. Furthermore, they were able to explain the DIF through differences in candidate intentions, college-bound versus non-college-bound (Ryan and Bachman 1992:21-22). Although Ryan and Bachman did not draw conclusions on the validity of the TOEFL reading section, the implication seems to be that since it is intended to be a test of academic English, their findings at least do not challenge the validity of the test. In terms of the kind of construct that Ryan and Bachman assumed, their finding of a motivational factor to explain differential item functioning indicates an interactional orientation.

6.6 Case summary

I will summarise the report on the TOEFL Reading case by answering the questions given in the case study protocol. These concern test development, validation, verbal construct definition, and the values that appeared to guide the development and validation of the test.

The procedures of test development reported in the TOEFL Reading case were detailed and highly structured. The draft testlets that had been written according to specifications were received by a test developer at ETS, they were reviewed and revised in a well-defined set of procedures that included peer review, sensitivity review and layout review, and then trialled. After data analysis, adjustments to the items were made in the light of the results and some items might be rejected. Content and psychometric descriptors were associated with the items, and they were saved in an item bank for operational use. The construction of operational test forms entailed

the use of statistical and content criteria to form a comprehensive test. After administration, the data were analysed and score reports produced. Within the examination system, the technical quality of the operational form was analysed and, as one of the monitoring procedures, operational item characteristics were compared with corresponding pretest data. The performance data were saved at ETS for further analyses of test quality and test constructs.

Compared with the summary framework of test development and validation in Figure 3, the model used by the ETS to develop the TOEFL Reading test followed the overall pattern. If anything, the case report highlighted the challenge and complexity of the work related to innocent phrases like vetting, editing, and pretest construction.

The criteria that seemed to guide the test development procedures were reliability, test form consistency, economy, professional accountability, and desire to understand what was measured. Psychometric quality criteria were clearly evident in all the studies published; they were especially evident in the monitoring and maintenance procedures that provided topics for some TOEFL research and technical reports. Professional accountability was demonstrated for instance through the availability of revision-by-revision data on a test form, so that Peirce was able to conduct her case study. Economy was demonstrated in the questions posed in several technical reports. In the test development and test form construction processes, psychometric criteria were combined with the content expertise of test developers. All in all, the efficiency and care with which the TOEFL Reading section was developed served the creation of high technical quality.

The range of studies published on the validation of the TOEFL Reading section shows that validity is an important consideration to the test developers. The ETS magazine *The Researcher* identified four areas of study under validity: construct, content, concurrent, and response validity. The Reading-related studies grouped under these categories showed a concentration on test scores. The construct validity question of what the test measures was asked, but this was often done from the perspective of the existing measurement scale, and thus the aim in the studies was to explain what the scores reflect in a measurement sense. However, some of the studies also concentrated on conceptual analyses and verbal descriptions of what was assessed. In addition to scores, these studies used conceptual categories from test specifications and textual and content analysis of items to study the nature of the tasks. A frequent problem expressed or implied in the studies was the lack of means to make a direct

empirical link from the considerations of test content to the test construct. The results of the score analysis showed repeatedly that variations in scores and score relationships did not reflect the content considerations. The content interpretation of the scores only proved successful indirectly through score relationships between different test sections or with other measures that include proficiency in English as one component. The conceptual information gleaned from these studies was that the TOEFL tested language proficiency and that the listening test seemed to measure a separable construct. More detailed content interpretation of TOEFL scores was not aimed at in TOEFL validation studies. The meaning of the scores was quantitative: the examinees were considered to have more or less of the ability indicated, and validation studies examined the measurement qualities of the test.

In relation to the areas of ongoing empirical validation identified in Figure 3 in Chapter 5, the validation research published on TOEFL Reading was narrowly focused on score properties and some aspects of score meaning. Some inquiries focused on score meaning and some, such as the speededness studies, might be categorised as tests of rival hypotheses. Revisions might have been proposed on the basis of some of the studies, but it is likely that these discussions were conducted internally among the members of the TOEFL Board or the TOEFL Committee of Examiners rather than in publications. Areas of validation inquiry listed in Figure 3 but not addressed in the studies discussed above include appropriacy of proposed test use, investigation of impact, and largely the examination of values that guide the development of the examination. Appropriacy of proposed use was presumably not addressed because the basic use for which the TOEFL is intended is well established, although another explanation for the absence of these studies in the present case is that it concerned a section of the TOEFL, not the whole test. Impact figured in the discussion once when the REM HERE and not far to go so go!

ETS (1999a:15) defined the construct measured in Section 3 as "ability to read and understand short passages that are similar in topic and style to those that students are likely to encounter in North American universities and colleges." The *Information Bulletin* for the paper-based TOEFL further specified that the section contained reading passages and questions about the passages, and that the examinees should reply on the basis of what was stated or implied in the passage (ETS 1999a:15). The *TOEFL Test and Score manual* explained that the items may focus on factual information presented in the passages, but examinees could also be asked to make inferences or recognise analogies (ETS 1997:12). In addition

to this characterisation, the *Information Bulletin* (ETS 1999a:15-16) demonstrated what the test was like through a practice passage and its associated questions.

The construct definition of the TOEFL Reading section quoted above can be considered interactionalist in the sense that the examinee's ability is contextualised in the academic environment. The material to be understood is text passages and questions, and the examinees are expected to understand information stated or implied. The definition is general, which is probably appropriate for a test taker audience. It is likely that the TOEFL test specifications contain a specific construct definition or at least define categories that test items must cover, but these have not been published. It is therefore not possible to analyse the nature of the definition in detail. Neither is it clear who developed the definition and when, but it is possible that it has been fairly similar since the introduction of the test.

The brief verbal construct definition was operationalized in test development and validation through factor analyses and studies of score relationships with other measures. Construct-oriented validation studies of the detailed test construct have not been published. It is possible that this is so because the paper-based TOEFL test is in its operational phase when the construct definition has been fixed. Unless a clear need for revision is detected, the activities are aimed at maintaining the current construct and keeping different test forms comparable in terms of the construct measured rather than improving the definition to develop the test.

Judging from the studies analysed in this chapter, reading is one of the skills that the developers of the paper-based TOEFL test consider necessary to test. A separate score for reading is reported even if the factor analyses on scores do not always indicate that this skill could be considered a separate measurement element in the test. The reading score included vocabulary items, which indicates that the test developers consider vocabulary and reading to be closely related, while the embedding of the vocabulary items in reading passages indicates that the test developers consider it important to test vocabulary in context. The items test comprehension of main ideas, details, and inferred information, but these were not found to constitute separable measurement dimensions in the test.

The item formats and assessment procedures selected for the paper-based TOEFL Reading test were based on multiple choice. Two obvious values served by this are reliability of scoring and practicality, given that hundreds of thousands TOEFL test performances must be scored each year. Test development practices made use of expert evaluation and psychometric criteria, and procedures of test construction indicated careful

attention to measurement properties. Thus, the delivery of accurate scores through carefully constructed tests was a high value in the TOEFL Reading case. When operational test forms were constructed, the measurement properties of the items and the whole test were consulted, and after the form had been used, the performance of the test was analysed and evaluated. Factor analytic validation studies and studies of score relationships were conducted, and suspicious test qualities such as speededness and bias were investigated. The activities of test development and validation supported each other especially where measurement quality was concerned.

Measurement information was not the only criterion used in the development and validation of TOEFL Reading tests, however. Theoretical acceptability was also observed, and analyses of test and item content indicated that the test developers were interested in knowing what their test measured. These considerations were combined with indicators of measurement quality, which implied that trade-offs were possible if content and construct rationales were strong enough to support them. However, the validation studies that analysed the content nature of the TOEFL test (Duran et al. 1985, Bachman et al. 1995) made it plain that the analysis concerned one form of the test, not the test specifications. Duran et al. (1985:64) proposed that further steps in content analysis could include the evaluation of the test's specifications against content analyses of test forms. The second proposal, to describe TOEFL items using communicative approaches "to understand how the content characteristics of TOEFL items are related to examinees' performance on items" (Duran et al. 1985:64) spelled out a design where the theoretical definition would guide test development and validation on a par with quantitative quality indicators, but to my knowledge such research has not been conducted with the paper-based TOEFL Reading test. Test development decisions were thus justified by content and measurement qualities with emphasis on measurement.

7 EXTENDED THEORETICAL DEFINITION OF CONSTRUCT: IELTS

7.1 Introduction to the IELTS case

The International English Language Testing System (IELTS) is a four-skills test aimed at assessing "whether candidates are ready to study or train in the medium of English" (UCLES 1999: inside cover), ie. a purpose very similar to the TOEFL test. IELTS is jointly managed by the University of Cambridge Local Examinations Syndicate (UCLES), The British Council, and the International Development Program of Australian Universities and Colleges: IELTS Australia.

The development of IELTS was initiated in 1986 and the test was introduced in 1989, at which point it replaced its predecessor, the English Language Testing Service (ELTS) test. After its introduction, the IELTS test has been modified once; the new format was introduced in April 1995. IELTS can currently be taken in 105 countries around the world. Approved test centres arrange the test on demand rather than on set dates, but usually at least once a month (UCLES 1999).

7.1.1 Boundaries of the IELTS case

This case focuses on the development of the IELTS test from when the development was first started to the present day. In terms of time, the present case thus begins from 1986 and covers 15 years. In terms of stages of test development, it encompasses both initial and operational development. All the sections of the test are included in the analysis because test developer reports exist on all of them. A computer-based version of the test is being developed (UCLES 1999:4), but its development is not included in the present case because reports on it have not been published.

The IELTS case represents the category in my case study design where the theoretical construct definition is extended while the psychometric definition is less extensive. This characterisation is motivated by my knowledge of the literature related to the test and by shared knowledge in the language testing world. Spolsky (1995:337-338), for instance, characterises TOEFL as psychometric and Cambridge Certificate examinations as humanistic and unconcerned with the importance of errors of measurement, but at the same time he notes that practices in Cambridge may be changing especially with the introduction of testing professionals on

the staff and in response to criticisms of poor measurement quality. Accordingly, the *IELTS Annual Review 1997/8* (UCLES no date:8-10) reports mean reliabilities of test forms and descriptive statistics for examinees from different backgrounds, for instance. Thus it cannot be said that psychometric considerations were neglected in the development of IELTS, but as the case will show, the theoretical construct definition had a significant role in the development.

7.1.2 Format of the IELTS test

The IELTS test assesses the four skills of reading, writing, listening and speaking, each in its own section or module, as they are called in *The IELTS Handbook* (UCLES 1999:4). Currently, there are two registration categories in IELTS, Academic or General Training. When IELTS was first introduced in 1989, there were three parallel academic modules; the change in 1995 combined these to the general Academic module. The new version is described here because the three Academic modules did not differ in their format but were different in their content. The listening module is the same for all candidates and the speaking module follows the same format for everyone. The reading and writing modules differ in content for the Academic and General Training categories, but the test length and the number of questions is the same. According to the *Handbook* (UCLES 1999:3), the Academic reading and writing modules "assess whether a candidate is ready to study or train in the medium of English at an undergraduate or postgraduate level", whereas the General Training modules emphasize "basic survival skills in a broad social and educational context" suitable for "candidates who are going to English speaking countries to complete their Secondary education, to undertake work experience or training programs not at degree level, or for immigration purposes to Australia and New Zealand."

The IELTS listening module is 30 minutes long and consists of four sections with altogether forty items. There are a range of possible item types: multiple choice; short answer questions; sentence completion; completion of notes, summaries, diagrams, flow charts or tables; labelling of diagrams; classification; and matching (UCLES 1999:5). Two of the sections deal with social situations and two with educational and training contexts. The reading module is 60 minutes long with three passages and altogether 40 questions. The item types are similar to listening except that instead of diagram labelling, headings are chosen for sections of text. Additionally, some items focus on identification of writer's views, attitudes or claims (UCLES 1999:6-8). The texts in the Academic alternative are

selected for prospective undergraduate and postgraduate students, and at least one text contains a detailed logical argument. The texts in the General Training reading module are characterised as factual informative, descriptive, and instructive, rather than argumentative (UCLES 1999:8).

The IELTS writing module is 60 minutes long and contains two writing tasks. One of the responses is required to be 150 words long and the other 250 words. The skills tested in each task are described in the *Handbook* in some detail. In task 1 in the Academic module, the candidates must present information from a diagram or a table. They are assessed on their ability to "organise, present and possibly compare data, describe the stages of a process or a procedure, describe an object or event or sequence of events, [and] explain how something works" (UCLES 1999:10). Part of the task realisation, according to the *Handbook*, is to respond appropriately in terms of register, rhetorical organisation, style and content. Task 1 in the General Training module involves writing a letter that requests information or explains a situation. The candidates are assessed on their ability to "engage in personal correspondence, elicit and provide general factual information, express needs, wants, likes and dislikes, [and] express opinions" (UCLES 1999:11). The language criteria with respect to register and so on are defined in exactly the same way as for the academic module. Task 2 is fairly similar in both modules, the candidates "are presented with a point of view or argument or problem" and they have to write a reasoned response. The topics differ according to the target group as in the Reading module. Both types of candidates are assessed on their ability to present a solution to a problem, present and justify an opinion, and present, evaluate, and challenge ideas. Academic candidates may also be evaluated on ability to compare and contrast evidence, opinions, and implications whereas General Training candidates may be evaluated on their ability to provide general factual information. (UCLES 1999:10-11.)

The Speaking module is "an oral interview, a conversation, between the candidate and an examiner" (UCLES 1999:12). It lasts 10 to 15 minutes. There are five sections: Introduction, Extended Discourse, Elicitation, where the candidate elicits information of the examiner, Speculation and Attitudes, and Conclusion. The section assesses "whether candidates have the necessary knowledge and skills to communicate effectively with native speakers of English" (UCLES 1999:12). Assessment takes into account "evidence of communicative strategies, and appropriate and flexible use of grammar and vocabulary" (UCLES 1999:12).

7.1.3 Developers of the IELTS test

The developers of the IELTS test include a project which developed the initial version of the test over a period of three years and an administrative board who funded the development and took over the responsibility of development and validation once initial development was complete. Spolsky (1995:337) explains that until the 1990s it was standard UCLES policy to be the administrative centre for testing activities but to acquire academic expertise for test development from outside. The policy has changed since, Spolsky notes, but when the IELTS test was developed, the project approach was still the formula used.

The ELTS Revision Project was directed by J. Charles Alderson, and the project membership spanned Britain and Australia (Alderson 1993:203). Alderson (1988:224-225) stated that for important parts of the work, the project was divided into seven teams which were devoted to the development of draft tests and specifications. The teams reported on their work in *Research Report* 3 (Clapham and Alderson (eds.) 1997) and their reports mention interaction between the seven teams and comments from a generic Project Team or a Project Steering Committee.

The formal testing board that has administrative control over the IELTS test was built on previous administrative structures. The IELTS predecessor, the ELTS test, was managed by UCLES and the British Council. When IELTS development was initiated, the board gained a new partner as the International Development Program of Australian Universities and Colleges (IDP Australia) joined the team. This broadened the range of potential users of the new test, and introduced the *I* for "International" in the name of the future test. Such broadening had been heralded by administrator perspectives on the evaluation of the old ELTS test, recorded in the ELTS validation project report (Hughes, Porter and Weir (eds.) 1988). Representing UCLES, Foulkes (1988:96) expressed the need to widen the user group of the new examination beyond UK universities. Initial administrative discussions might already have been in progress at that stage. IDP Australia has subsequently formed a division called IELTS Australia. In addition to its managerial activities, this division funds research into IELTS in use in collaboration with the British Council and UCLES (Wood (ed.) 1998, Foreword).

7.1.4 Test development brief: conditions and constraints

IELTS was developed to replace the ELTS test, an English for Specific Purposes test that had been introduced in 1980 to provide British universities and other institutions with a test of English language proficiency

available on demand world-wide (Westaway, Alderson and Clapham 1990:239). The test was based on Munby's (1978) specific purposes model, and in the course of a few years it began to attract criticism that gave the examination board cause for concern (Spolsky 1995:344). It was said that the test was administratively cumbersome and that it was based on an outdated model of language ability (Alderson 1988:224, Alderson and Clapham 1992:150, Alderson and Clapham (eds.) 1992:2). However, as Alderson and Clapham (eds.) (1992:2) point out, the project team were advised that the revision should not be too radical because the users of ELTS were generally happy with the test. Criper and Davies's ELTS validation study (1988:22-25; 81-89; 108-109) provided clear evidence for this: admissions officers reported no problems in using ELTS scores; language tutors approved of the test in general although they also had some questions and criticisms; two thirds of the test takers found the test fair or quite fair, and almost nine tenths felt that the test reflected their proficiency accurately. The revision project's brief was to make the revision "in the light of operational experience over 10 years, feedback from test users, and the results of the Edinburgh ELTS validation project" (Alderson 1993: 203).

In practice, the request for continuity meant that some central features of the ELTS test and its reported scores had to be carried over into the new test. Specifically, Alderson and Clapham (1992:154) indicated, the 9-point band scale for reporting scores should be maintained, and scores should be reported in a profile, which meant reporting separate scores for reading, writing, listening, and speaking. Furthermore, the ELTS test had contained subject-specific modules as well as generic sections as the test takers had found these appealing. The project brief was therefore that some modularisation should remain. At the same time, however, "the test should be shorter, administration should be simpler, and the revised test should be more reliable. Because of financial constraints, the speaking and writing subtests should be only single marked, and all other subtests should be clerically markable" (Alderson and Clapham (eds.) 1992:2).

The most specific guidelines of the development brief concerned score reporting. The project team was also instructed to begin the development with a thorough consultation of all the stakeholders involved in the current ELTS and the future IELTS test.

Once the IELTS test was published, the activities began to be controlled by the test specifications and the administrative procedures which had been set at publication. These have not been published in their entirety, but those parts of them that are included in the reports and articles will be discussed in the course of the analysis below.

7.2 Nature and focus of studies published on the IELTS test

In this section, I will give a brief initial analysis of the studies that have been published on the development and validation of the IELTS test. I will discuss them in two groups, those that concern initial development and those that have been published on the operational development and validation of the test. Since the literature on IELTS is less extensive than that on the TOEFL, the treatment is quite concise.

Several reports on the initial development and validation of IELTS have been published. Some of them were brought out by UCLES, while others are articles in academic books and journals. The initial development of IELTS also provided material for at least one PhD thesis (Clapham 1996a). All the publications were written by the people who were actively involved in developing the test.

A summary table that gives detailed information about the reports and studies on the initial development of IELTS can be found in Appendix 3. The table identifies the works by their author and date of publication and also specifies the focus of each study, the materials and methods used, and the main findings. The last column in the tables indicates what issues in test development and validation each study concerns. The categorisation is mine, since the testing board has not published their own categorisation.

The nature of the studies on the initial development of IELTS is very different from that of the TOEFL. Instead of following the basic outline of empirical research papers where a statement of the research problem is followed by an account of materials and methods, after which comes a presentation of the results, and the report is closed with a brief discussion, the IELTS initial reports focus on the procedural nature of examination development. One of the inputs is always the writer's participant knowledge of the development process, and the papers discuss the ways in which the developers made use of different strategies and sources of information when they developed the test. Instead of reporting "findings", the studies constitute records of development rationales and decisions, a practice that is recommended in the current *Standards for Educational and Psychological Testing* (AERA 1999). An exception in this regard is Clapham's (eg. 1993, 1996a) study of the effect of background knowledge on reading comprehension, where the researcher carefully defined research questions and implemented a number of designs to look into them. The publications on the initial development will be discussed later in this chapter.

The post-publication research reports on the IELTS test are summarised in Appendix 4. Similarly to Appendix 3, the reports are

identified by their author and date and summarised in terms of focus, materials and methods, main findings, and concerns of test development and validation that they address. Some of these reports were written by "insiders" such as test developers, assessors or interviewers while others were written by researchers external to the development team.

The reports on the development and validation of IELTS during the operational stage are structured more traditionally than the reports on initial development. They tend to have formally defined research questions and accounts of materials and methods used, followed by careful reporting of the results. At the same time, these studies are less closely related to the actual test development activities than the reports on the initial development. Only the implications of the studies from the operational stage of IELTS are directly concerned with possibilities for development in the examination. The studies address a wide range of concerns from improving the quality of testing procedures through score comparability to authenticity, impact, acceptability, and score use. As far as the validation results are concerned, these could be used in building validity cases for score use in other similar settings. However, it is evident from many of the studies on operational IELTS that they have not been written by people who are directly concerned with the development of the test.

7.3 The starting point for IELTS development

To contextualise the IELTS case, a brief summary must be given of two key documents on the test's predecessor, namely the *ELTS Validation Project Report* (Criper and Davies 1988) and the proceedings of a conference that UCLES held to consider the implications of the validation report (Hughes, Porter and Weir 1988). While these two reports are not directly part of the IELTS development process, they give important background information for it.

Criper and Davies (1988) conducted a validation study of the ELTS test. Their aims were to examine the concurrent validity of ELTS against two other existing tests and against success in academic studies, to examine the predictive validity of ELTS in relation to students' success in academic studies, to examine the reliability of the test, and to examine its face, content and construct validity (Criper and Davies 1988:13). Samples of 187 to 195 students took ELTS and the two other tests with new samples on each of three years, and to provide concurrent data on their performance, a self-assessment questionnaire was administered and judgements were gathered from their supervisors and language tutors. The students also answered a

detailed background questionnaire. Reliability was assessed through internal consistency and through test-retest measures with a one-month and a nine-month interval. Face validity was assessed through student questionnaires and content and construct validity through a careful content analysis. In addition, questionnaires were administered to admissions staff at universities. The methods used to analyse the data included descriptive statistics, correlation, multiple regression, and factor analysis, the latter concerning ELTS section scores. Criper and Davies (1988:50-52, 56-57) found that ELTS total scores correlated quite well, from .77 to .85, with those of other batteries but not well with tutor ratings. Similarly, the prediction of overall academic success was .30, which was comparable to the results achieved with other proficiency batteries (Criper and Davies 1988:63-76). The students found the test acceptable and the administrators found it interesting but long and cumbersome to administer. The multiple choice sections were found to be quite reliable (internal consistencies ranged from .80 to .93), but the writing and speaking tests were found worryingly unreliable at about .50. The conclusion of the evaluation was that "in its own terms [ELTS] is a satisfactory test of English proficiency because of its reliability and certain claims of validity" (Criper and Davies 1988:114). However, the both the predictive validity and the practicality evidence suggested that a shorter and more easily administered test would be more desirable.

The papers from the seminar that was arranged to evaluate Criper and Davies's report in 1986 concentrated on construct, content, concurrent, and predictive validity and practicality. The issues raised included the serious difficulties with validation in the absence of test specifications, concern with low reliabilities, and some implementation problems with the validation study such as the low numbers of examinees in some of the validation samples. The reports on construct validity all raised the need to define the complex construct of the ELTS in more detail (Weir et al. 1988:9). The question was also raised of the conflict between the complex conceptual definition of the construct and the apparent conflict that the test seemed unidimensional in a measurement sense. There was also the concern that the low reliability of some sections made it difficult to argue that any construct was consistently measured in the test at all (Weir et al. 1988:9, Henning 1988:87). The overall feeling was that the general acceptance of the test favoured its revision rather than replacement, thus justifying the ELTS Revision Project.

7.4 Initial development of IELTS

7.4.1 Stages of IELTS development

In accordance with the project brief, the development of IELTS began with an evaluation of the existing ELTS test and an extensive data gathering exercise on the views of ELTS users (reported in Alderson and Clapham (eds.) 1992 and Westaway, Alderson and Clapham 1990). The purpose of this was to establish the starting point and specify exactly why and how ELTS users thought the test should be changed (Alderson 1988:224).

Alderson (1988) lists five further stages in the development of IELTS. Stage 2 comprised a setting up of seven project teams to produce draft specifications and tests. Alderson (1988:225) points out that the teams were asked to develop both at the same time because they would have to operationalize the specifications at some point in any case, and because coherence between the two was important for validation. At stage 3, reactions to draft items and specifications were gathered from those who should know about candidates' language needs, i.e. subject specialists, pre- and in-sessional language teachers, applied linguists, testers, and students. Alderson (1988:220, 225) suggests that this constitutes an innovation in validation techniques, especially concerning content validation, in that reactions were gathered to both the specifications and their operationalization into draft items. Stage 4 consisted of the preparation of final specifications and modification of the sample items according to feedback, and the production of trial tests on the basis of these documents. At stage 5, the sample tests were tried out and predictive validity data were gathered, and at stage 6, the final forms of the test, training manuals, and practice materials were produced. (Alderson 1988:226.)

The IELTS way of structuring and reporting test development emphasizes the very first steps in the test development process, mainly a detailed analysis of the rationale and goals and the development and validation of test specifications and draft tasks. In all of these stages, stakeholder comments were used in addition to the developers' views. Further analyses of student needs were not conducted because such analyses already existed and because stakeholder views were required to complement the developers' knowledge of the social need for the test (Alderson 1988:222-224).

Compared with the generic framework of test development presented in Chapter 5, the only point that did not seem to receive much emphasis in the overview of the development of IELTS was the forward-planning part of

the initial validation work. That is, the IELTS development overview does not refer to the task of identifying plausible rival hypotheses for the proposed score interpretations or the planning of future validation research. Alderson (1988) mentions validation in the discussion concerning specifications, which indicates that the developers considered validation to be a part of test development. The reported focus of the work was the recording of justifications for test development decisions. Forward-planning discussions may have been conducted, but they were not recorded in the articles and reports published.

7.4.2 Work on construct definition

The IELTS development brief included a request to revise the "outdated" construct of the ELTS test (Alderson and Clapham 1992:150). Accordingly, the developers reported on construct definition work in considerable detail.

The project started from a divisible construct of language proficiency which had been operationalized in the existing ELTS test (Criper and Davies 1988:9-10). The construct was divided into components in several ways. Firstly, there was a division between the four skills of reading, writing, listening, and speaking. ELTS had separate tests for each skill and the scores were reported in a profile. Secondly, a distinction was made in ELTS between "general" and "study" skills such that study skills were tested in a separate section. Thirdly, proficiency was divided according to subject matter specialism, ie. ELTS was a test of English for Specific Purposes (ESP), based on the Munbyan idea that language use was first and foremost based on the needs of individual language users. There were six different academic modules: life sciences, medicine, physical sciences, social studies, technology, and a "general academic" module for other areas of academic study. There was also a generic "non-academic" alternative for the testing of intending trainees whatever their specialisation. Criper and Davies (1988:6) pointed out that the selection of modules "created and creates numerous problems and difficulties and raises, in an extreme form, the debate about the multi-factorial/uni-factorial structure of language tests and of language abilities." The ELTS validation study concluded that the modular division was not effective (Criper and Davies 1988:114).

Alderson and Clapham (1992) described the first steps in the drafting of a new construct definition for IELTS. The project decided to conduct a survey of recent views on the nature of language proficiency to assess whether there were any strong candidates to replace the ELTS construct. They sent a letter to 22 applied linguists in the United Kingdom and North America which outlined the aims and constraints of the test development

project and invited the specialists to express their views on an appropriate model of language proficiency for IELTS. They received 11 replies. The main finding was that there was no generally accepted dominant paradigm or theory of language ability which the respondents could have suggested (Alderson and Clapham 1992:155). A relatively unanimous recommendation was that the test developers should look into kinds of language use that the test takers have in common as well as those which differ. The applied linguists found the idea of a division into the four skills of reading, writing, listening and speaking fairly acceptable. They also found it important that the test should somehow reflect 'real life' language performances (Alderson and Clapham 1992:162). Other recommendations varied from individual to individual.

Alderson and Clapham (1992:164-165) concluded that the only alternative that the test development project had in the face of this lack of clear models was to be eclectic. They suspected this could have been the result in the late 1970s as well if a similar investigation had been conducted when ELTS was first developed. They claimed, however, that eclectic models that test developers implement can be useful (Alderson and Clapham 1992:165). These take into account theories of language ability but also "variables associated with test purpose, audience, and the practicalities of test design and administration which are not the concern of theoretical applied linguists". Such models, they stated, can contribute to applied linguistic theory when testers examine the way in which their tests work.

Thus Alderson and Clapham seem to be making a case for what might be described as inductive construct definition. The test developers deduce the properties of the construct that their test implements through their insider knowledge of the test purpose, the test specifications, the draft tasks and criteria used in their revision, the drafts and re-drafts of the assessment scales, and the developers' perceptions of the skill that they intend their tests to focus on. Recording this definition in the test specifications, developing hypotheses about relationships between different scores, and designing and conducting studies on the test scores might provide interesting construct information for both construct validation related to the test and for applied linguists working in other areas than language testing.

Unfortunately for the purposes of the present study, Alderson and Clapham (1992) only discussed the advice given by the applied linguists and other stakeholders. They did not continue to describe the construct definition which the board may have drafted and which they implemented in the new IELTS test. However, the same authors did characterise the nature

of the operationalized construct in broad terms in *IELTS Research Report* 2 (Alderson and Clapham (eds.) 1992), in connection with an account of the process by which the content and structure of IELTS was developed.

The decision was made to have both a general section and a modular section in the new test. The construct of the new test would thus continue to be divisible and partially subject-specific. The general section would be the same to all test takers, and it would contain tests of lexis and structure, listening, and oral interaction. According to the developers, this decision represented a compromise between desirable construct definition and practical testing solutions. *Research Report* 2 (Alderson and Clapham (eds.) 1992:15-18) stated that good arguments had been presented for the inclusion of all four skills in both the generic and the subject-specific components, but this could not be done because the test had to avoid duplication to be short and practical. Listening was included in the general component even though the needs analyses had indicated that one of the main challenges for non-English-speaking students was to understand lectures. This was done for the practical reason that most test centres were not able to arrange separate listening tests for three groups of test takers. Oral interaction was to be tested in the general part of the test because specific-purpose interactions had proved artificial and difficult to conduct during the ELTS test. The interlocutors could not possibly be experts in all the topic areas of their candidates, so expert-expert discussions were difficult to simulate. The reading section, which had been part of the general component in the old ELTS test, was moved into the modular component in the new IELTS.

Alderson and Clapham ((eds.) 1992:10-15; 17-18) recount how the subject specificity aspect of the IELTS construct evolved. After careful consideration, the project made a decision on a division into three academic specialisms: business and social science; physical sciences and technology; and life and medical sciences. This division had been one of the suggestions made by the specialists consulted, and it was confirmed after a sample of existing score reports indicated that a division into these three groups would divide the IELTS population into three equal-sized groups. Initially, the project did not make decisions beyond this division, but later, a decision was made to develop a separate modular section for students in vocational training.

Thus, the new construct maintained two of the old construct's divisions. The division into the 'four skills' remained, and there were three subject-specific alternatives within the test. Furthermore, a generic alternative was developed for test takers who did not intend to go in for

academic studies. Study skills were blended into all the tasks of the test and were no longer assessed as a separate construct.

Some overall evaluations of the construct(s) assessed in IELTS were presented by the test developers when they discussed the development and revision of the test specifications and tasks (Clapham and Alderson (eds.) 1997). An example is Foulkes's (1997) discussion of the listening module. He stated that the working group's initial definition was on a very general level; they decided that both competence and performance should be tested in the listening section (Foulkes 1997:3). The second version of the specifications characterised the construct in more detail, stating that the test assessed "the candidates' ability to perform a range of tasks such as 'following and responding to instructions', and 'retrieving general factual information' within a set of topics such as 'travel', 'accommodation', 'recreation' and 'education'" (p. 4). According to Foulkes (1997:5), the team had intended the test to contain plausible spoken language with hesitations, self-corrections, and shifts of register, but the attempt had been partially defeated in practice, because all the listening material was scripted before recording. In this case, too, practicality overrode theoretical desirability.

Foulkes (1997:12) concluded his account of the development of the IELTS listening section with a rather critical self-evaluation: Judging by the test's high correlation with the old ELTS listening test and the new IELTS grammar test, the test developers were "too successful" in developing a test of general listening ability, and in fact produced a test of general proficiency. At least in construct terms, the team could not be sure what the construct was that they were measuring, albeit reliably.

In the future, analyses of the construct implemented in the testing process could be conducted for instance along the lines described by McNamara (1996) or Buck and Tatsuoka (1998), as discussed in Chapter 4. Combined with the test developer perspective, such an approach might offer the kinds of insight Foulkes (1997) seems to miss.

The most detailed descriptions of the construct(s) assessed in IELTS are presented in the appendices to *IELTS Research Report* 3 in extracts from the specifications for listening, grammar, reading, writing, and the general training module (Clapham and Alderson (eds.) 1997:125-163). The extracts provide fairly clear examples for one of the two purposes of test specifications identified by Alderson (1988:228-229), that of providing detailed guidelines for task writers. However, the extract from the writing specifications also presents a description of the test content and construct, which is the second purpose that Alderson (1988) identifies. For the present

purposes, this is useful because it describes *what* is assessed, as well as how to develop sets of tasks to assess this.

The writing specifications defined test focus in terms of targeted band levels, academic tasks to be tested (eg. "organising and presenting data", "explaining how something works", "arguing a case"), and appropriate audiences (Clapham and Alderson (eds.) 1997:146-147). This selection of features shows that the test developers considered it important to contextualise language use in the test tasks to the degree that it is possible in a test. The candidate has the role of an academic student and the tasks broadly simulate the challenges of academic writing, at least those of them that can be simulated in a short time in a test setting. The intended audiences are professorial, professional, and personal. The broad definition of task difficulty shows that the construct definition has been written for test purposes: the concept of difficulty offers task writers a shorthand for describing one aspect of limited-duration, assessment-oriented language tasks.

After the generic construct description, the writing specifications include more concrete characterisations of tasks: what the stimulus texts can be like, how the prompts should be written, and what the organisation of the tasks within the writing section is. This is followed by detailed definitions of test tasks and a template for producing parallel tasks (Clapham and Alderson (eds.) 1997:147-153). This type of detailed construct definition affords a very good basis for research on the relationship between the construct definition and the procedural construct implemented while test takers are taking the test, as well as the relationship between test and non-test writing by examining the nature of tasks and learners' performances on them. Similarly, features of test tasks could be manipulated systematically and results investigated. Such research can only be conducted on the basis of specially set research designs, but they might be linked to operational test data by using some tasks in exactly the same form as they appear on an operational version of the test.

7.4.3 Development of the specifications and tasks

Evaluation of initial test designs and content validation were conducted for IELTS by showing initial specifications and draft tasks to a range of stakeholders and soliciting comments (Alderson 1988, Clapham and Alderson (eds.) 1997). The comments resulted in changes both in the specifications and in the tasks. Most of these appear to have led to further gradual specification and clarification as explained in Foulkes's (1997) report on the speaking specifications and test discussed above. All the team

reports in *IELTS Research Report* 3 made reference to skill definitions and guidelines for task writing in the specifications (Clapham and Alderson (eds.) 1997).

Judging by the developers' reports, the most visible changes to the specifications and tasks based on expert comments concerned the three academic reading and writing modules (Clapham 1997, Hamp-Lyons and Clapham 1997). The initial specifications and tasks for each of the three modules had been different, but as a result of the content validation round they became fairly identical. Expert comments influenced this decision, but so did practicality considerations. One team had proposed a relatively large set of readings with the sole task of writing an essay based on those of them which were relevant. While the commentators found the task interesting and fairly authentic, it was unconventional, and made it difficult to give separate grades for reading and writing, which was the project brief. A more traditional line was therefore adopted, with separate tests for reading and writing, and with the skills being defined in terms of academic tasks such as 'identifying underlying theme or concept' and 'arguing a case'. The report indicates that, after this stage, the specifications were changed gradually over time, and the final changes were only made after the results of the pilot and trial tests had been analysed (Clapham 1997:57).

7.4.4 Development of IELTS assessment criteria

IELTS Research Report 3 includes a reprinted paper on the development of the band scales on which the IELTS scores are reported (Alderson 1991). As with the test instrument, the project started from what was already there in the form of the ELTS scales. Their brief was to retain the 9-point reporting scale, but if possible, simplify its use.

In the paper, Alderson (1991:72-75) distinguished between three functions of scales. User-oriented scales reported the nature of the assessed performance to test users, assessor-oriented scales guided the assessment process, and constructor-oriented scales guided test development. The differences had become apparent when the IELTS project team attempted to create skill-specific reporting scales. The first drafts of the scales had included descriptors which did not correspond to what was being tested in the tasks, and it was not possible, nor had the project intended, to allow the scale development to influence test content. As it had proved difficult to create clearly worded and useful descriptive scales for reading and listening, the project decided to report the results of these tests on the overall scale only. The same reporting practice was adopted for the speaking and writing

tests, although for these tests, separate scales were developed for assessors to use while assessing performances (Alderson 1991:76).

The development of the assessment scales for speaking and writing followed a four-step procedure, of which the first step of initial drafting Alderson (1991:81-82) did not discuss. After a set of draft descriptors for the overall scale existed, they were sent to moderators, teachers, experienced markers, and other experts for comments. Many anomalies and inconsistencies were removed as a result. Next, experienced markers were asked to identify a set of sample scripts which typified each of the holistic scale levels, and agree on what the features were that characterised each level. The criteria were then listed and level descriptors were created and revised against the sample scripts in an iterative process. The third step involved feedback from scale users when they were assessing pilot performances and constructing assessment guides. Alderson argued (1991:81-82) that the result is inevitably a compromise driven by the usability of the scales. The ultimate criterion is practical and based on the demand that assessors must be able to agree approximately what the levels mean.

Ingram and Wylie (1997: 24-25) briefly summarised an attempt that the IELTS Speaking team made to introduce an assessment flowchart instead of a more traditional assessment scale. In the trial scale, the assessors were first asked to distinguish between three levels of intelligibility, then four levels of fluency, next five levels of accuracy and range of grammar and lexis, and finally give scores on a 9-level scale on appropriacy, functional range, initiative, and pronunciation. Each of the assessments was dependent on the level chosen at the previous stage, so that once a starting point was identified, the assessor only had to made a binary decision at any subsequent assessment point. However, the team found such a scale difficult to construct within the time constraints of the project, and the attempt was abandoned. The approach foreshadows that of Upshur and Turner (1999) discussed in Chapter 4, except that in the case of IELTS the scale was to be general rather than task-specific. The difficulty of finding generic criteria that allow successive refinement of assessment for all the different performance possibilities across different tasks may well explain the difficulties that Ingram and Wylie (1997) referred to.

In another article, Ingram and Wylie (1993) discussed the overall assessment scale that they decided to use in IELTS in contrast to 'functional' scales, which define what learners at each level 'can do'. The contrast that they identified was that the IELTS scale, like the ASLPR one, defined *how well* the learner was able to use the language in spoken

interaction rather than listing *what* they were able to do. They argued that a scale that described quality of linguistic performance was more helpful for assessment purposes than a purely functional scale because it helped distinguish between learners who can cope with similar functional tasks at different degrees of linguistic sophistication. They also argued that "nonfunctional" descriptors that characterise quality of language produced may help extrapolate the learner's proficiency beyond the particular tasks which the learner completed under test conditions (Ingram and Wylie 1993:221-222). The construct implication of this claim seems to be that there should be a generic dimension of proficiency that is not very strongly situation-related but rather influenced by an individual's language knowledge. This may be what the IELTS speaking scale attempts to achieve.

7.4.5 Pre-publication piloting and review and revision of test materials

The results and implications of the smaller-scale piloting and the larger-scale trialling of the draft IELTS tests were discussed in some detail in the various papers in *Research Report* 3. Some item statistics are provided for the reading, listening, and grammar tests, while the piloting experiences with the speaking and writing tests were not presented numerically. The most extended psychometric report was Griffin and Gillis's (1997) analysis of the objectively marked sections and a detailed example of test analysis with the Reading section of the Physical Sciences and Technology module. Their overall conclusion on the final, larger-scale trials across 17 countries was that the IELTS test provided reliable results which were interpretable in a stable manner in the different geographical and linguistic settings in which IELTS would be used (Griffin and Gillis 1997:123).

The reliabilities of the objectively marked IELTS sections ranged between .84 and .92. The sample sizes on which these were based ranged from 232 to 842, lending credibility to the results. The results of the IRT analysis of the PST Reading test indicated that the most accurate decisions, shown through the smallest Standard Errors of Measurement associated with the scores, were being made in the range of 4 to 6.5 on the IELTS scale (Griffin and Gillis 1997:116-117). This was appropriate because the most common cutoffs were likely to be made at this score range. However, Griffin and Gillis (1997:119) raised the concern that higher cut scores of 7 or 7.5 might be used for decisions on postgraduate study. This was worrying because there were no items at this level; thus the confidence that

could be placed on the decisions was limited. They recommended that items should be written for these difficulty levels.

Some changes and modifications were made to the IELTS test during the piloting/trialling stage. Most of these were minor, but some more extensive changes proved necessary with the writing and grammar sections.

Regarding writing, the smaller scale piloting did not indicate major problems, but the main trials revealed that some of the writing tasks were unsatisfactory (Hamp-Lyons and Clapham 1997:75). Some tasks included unrealistic features in terms of purpose of writing or target audience, in one the topic was too trivial, and in one the range of language elicited was too narrow. This showed that the specifications had not been detailed enough, so that test versions which had been intended as interchangeable were in fact somewhat different. This, in turn, led the project to refine the writing specifications.

Hamp-Lyons and Clapham's (1997:74-79) report on the development of the writing tasks, assessment criteria, and specifications illustrates the flexibility with which test development proceeds prior to the publication of the test. Every stage and every component has the potential to influence others. In their case, variability was found in the difficulty of the writing tasks during the main trials. This led to a revision of the specifications. A detailed construct definition was included, and task templates were inserted to guide the construction of equivalent tasks. The templates specified the content and format of the tasks, alternatives for the kinds of writing to be elicited in each task, and levels of difficulty. They also provided sample instructions for the tasks. Once the templates were created, the existing tasks were revised according to the new specifications, and where this was impossible, new tasks were written. This gave rise to further refinements in the specifications. The new tasks were then trialled, and in connection with trialling the use of the assessment scales and the construction of assessment guides, the tasks and specifications were adjusted further. During this cyclical process, the construct description, the instructions for its operationalization, and its actual realisation in tasks and scales were iteratively refined.

The grammar test was constructed and trialled in a way similar to the rest of the IELTS tests. The first draft tests and specifications were constructed in an iterative process. The tests focused on "a student's ability to process and produce appropriate and accurate forms in meaningful contexts" (Alderson and Clapham 1997:46). Since the tasks were based on texts which the candidates had to read in order to answer the items, the

project members tried to take great care to make the test foci for grammar and reading different.

The grammar specifications and tasks were revised in consultation with other project members, and the revised tests were trialled on fairly large samples of test takers. The results indicated that the test was fairly reliable, the KR-20 indices ranged from .82 to .91 (Alderson and Clapham 1997:37). However, the grammar test turned out to correlate well with the rest of the tests in the IELTS battery, and especially closely with reading and listening. This indicated that the tests assessed similar, possibly the same skills. As the test results were to be reported separately for the four skills, and as analyses indicated that removing the grammar test would not affect the reliability of the test battery adversely, the decision was made to drop the grammar test from the battery, thus saving 30 minutes of candidate testing time. The rationale for this decision was both construct-based and practical; a shorter test would serve the interests of all the parties, especially since no important information and no measurement qualities appeared to be lost. The way in which the skills tested in the deleted grammar section were content-wise or procedurally related to the other IELTS sections was not investigated as part of the IELTS development work.

7.4.6 Development of administrative procedures

The development of administrative procedures is not reported in the IELTS development literature in great detail. This may be because an administration infrastructure already existed for the predecessor, the ELTS test, or because the developers did not consider this aspect important to report on. Furthermore, overall concerns of practicality had been taken into account in the project brief. The only aspect of administrative procedures which required more extensive attention from the test development viewpoint was the assessment procedures, especially those for writing and speaking.

Alderson's (1991) report on the development of the IELTS assessment scales discussed the need for assessors to agree and the obligation of the examination system to monitor examiner performance, re-accredit them at regular intervals, and exclude those assessors who cannot conform to the shared view. Otherwise, comparability of scores would be endangered (Alderson 1991:81). This practical view of ensuring the quality and comparability of grades was reflected in the working groups' reports on the writing and speaking components of IELTS.

Hamp-Lyons and Clapham (1997:79) discussed the creation of a Writing Assessment Guide for IELTS assessors very briefly. They stated that the Guide included explanations of the marking criteria and sample

scripts which were marked and commented in terms of why the marks had been given. They also stated that "minor final amendments" were made to the marking descriptors as the guide was written. Furthermore, a training program for writing examiners was devised and a Certification Package was created. The criterion used in the training program was that prospective assessors had to show that they "were marking within acceptable limits" before being certified (Hamp-Lyons and Clapham (1997:79). The implementation of the training program was not reported.

Ingram and Wylie (1997:18-19; 25-26) reported briefly on the training procedures developed for the interlocutor-raters of the IELTS speaking section, and also described the quality control procedures from the examination's point of view. The aim of both the training and the quality control was to ensure comparability. Furthermore, the test format selected, structured interview, was intended to further support the comparability of the assessment procedure across testing events.

Two interlocking sets of materials were developed for IELTS oral assessment: a training package for examiner-assessors, and an introduction to the documentation of the test. The training involved an introduction to the principles of oral examination and guidelines for administration, observation of taped interviews with comments on salient features, and actual conduct and rating of practice interviews (Ingram and Wylie 1997:25-26). The documentation of the test, to be studied as part of the training package, contained an introduction to the speaking test, an administration manual, and an assessment guide. The team recommended that examiner-assessors should be required to go through the training and successfully complete practice interviews and assessments before accreditation, and that they should be required to re-visit the material at regular intervals and assess further performances to be re-accredited. A monitoring system whereby all interviews are taped and a random 10% of them re-rated centrally to observe quality was also created.

Ingram and Wylie's discussion of the training and quality control principles concentrated on the nature of the procedures and the way they were developed. At the end of their report, they listed fifteen points of interest for further research, including investigations of score reliability, studies of validity in terms of test and score comparability, the usefulness of the scale descriptors to raters and score consumers, and the effectiveness of the monitoring procedures for the reliability of IELTS. A list like this is a concrete example of commonalities in test development and validation. The questions on Ingram and Wylie's list focus on learning more about how the test works, both in itself and in comparison to others. A further

development of this research program would constitute a prime example of bottom-up construct work in test development.

7.4.7 Validation work

As has become apparent from the summaries above, validation was one of the strands of activity in the initial development of IELTS. It was not a separately labelled strand but rather a significant ingredient in many if not all of the development activities. It was particularly visible in the way the work on specifications and tasks was reported.

Alderson (1988:225-226) made the case that an iterative process of developing items and specifications and soliciting the views of stakeholders to inform further versions of both, as was done when IELTS was developed, could be viewed as validation. He also expressed the hope that this dynamic process would constitute a more practical procedure for test development ”than the standard ’needs – specifications – items’ model that seems so prevalent in much of ESP” (p. 229).

If content validation is traditionally seen to happen at the final stage of test construction and take the form of expert judgement, the two main differences implemented in the development of IELTS were that a broad group of stakeholders was included in the commenting process and that the consultation began early in the test development process to inform the course of the development. This included the development of the construct definition. The body of data produced during the consultation process could be used as early validation evidence. Further validation work could be planned on the basis of the data; for instance, the critical remarks made by the reviewers could be scanned for plausible rival hypotheses for explaining variation in test scores.

Hamp-Lyons and Clapham’s (1997) report on the development of the writing specifications and tasks showed that the writing team’s experience with trials of actual administration led them to revise the specifications, because the results showed that the earlier version had allowed too many different operationalizations. The revision enabled the team to tighten both the construct definition and the guidelines for operationalization. This resulted in a test which was more clearly-defined and controlled. At the same time, this process showed the degree to which validation-related work was necessary for test development. The validation work could be carried further by developing studies on operational test material to see whether the construct as defined in the specifications and assessment criteria is actually assessed in the testing process and whether it corresponds to writing in the academic study setting.

In the summary framework of the test development process, I categorised validation work during initial test development into four groups, which were recording of the processes and products of initial test development, evaluation of the products against aims and the state of theory, identification of plausible rival hypotheses, and planning of future validation research. All this may well have been done by the developers of IELTS but only some of it has been published. The development process was recorded and published in some detail, as was shown in the discussion above. It is likely that this led the developers to evaluate the process and products of the development activities, but the evaluation has not been published. The evaluation would likely lead to a program of test development, monitoring, and validation for the future, but if such a plan was developed, it has remained internal to the examination board. The *IELTS Annual Review* 1997/8 (UCLES no date:11) stated that "all IELTS research activities are co-ordinated as part of a coherent framework for research and validation", which indicates that such a plan exists. The proposals for further research in Ingram and Wylie's report and article (1993, 1997) may well be part of this plan. Some evidence to support this assumption is offered by the studies on IELTS given out after the publication of the test, some of which follow up Ingram and Wylie's proposals.

A significant strand of research related to the validation of IELTS is Clapham's study on the defensibility of the subject-specific reading tests (Clapham 1993, 1996a, 1996b, 1997). The main body of data came from IELTS trials, but the results of her research only came out after IELTS was operational. Her main result from the test development point of view was that students might benefit from reading texts in a discipline broadly in their own area, but they might also be disadvantaged if the text was highly specialised. If the latter happened, it would threaten to invalidate the results. Since a single-module test was also more simple administratively and since such a solution reduced costs, the decision was made to change the examination. From April 1995 onwards, IELTS only contained a single reading module, and also a single writing module, for all intending university students.

In her study of IELTS reading, Clapham (1993) started from the 1989 version of the IELTS reading test, which had three distinct academic specializations: business and social science, life and medical sciences, and physical science and technology. In a pilot study, Clapham found "no evidence … that students are disadvantaged if they take a reading module outside their academic discipline" (1993:270). The IELTS examiners might

thus cautiously draw the conclusion that a single academic module might be enough for the test system. However, as Clapham explained in detail, the study was restricted in scope, so the conclusions must be considered tentative. The number of cases investigated was small and the study operationalized background knowledge only through the students' self-reported field of study. Furthermore, Clapham (1993:267) made the point that the empirical results are not the only factor to inform test design, and acceptability to users also plays a role in design decisions. Acceptability had been the reason why the three specialist modules had been retained in IELTS.

Clapham (1996a, 1996b) delved further into the issue of subject specificity. With a larger number of subjects (N=203-328), the result was that, overall, students did better in their own subject area. When the results were analysed text by text, however, the differences were not always significant, and in one case the students did better on a text which was *not* in their area. Several of Clapham's results related to the concept of subject specificity, which she found complex and elusive of clear operational definition. From a test development point of view, Clapham (1996b:189) argued that the avoidance of negative effects was the main concern. In spite of acceptability to users, if it was possible that students might be disadvantaged by a text choice favouring candidates from other discipline areas, this critical concern outweighed all others in threatening to invalidate test results. Validity concerns thus argued for a single test for all.

In sum, the validation work on the initial development of IELTS comprised the recording of development rationales and justifications for the decisions made. Clapham's more formal validity design similarly focused on test development decisions, but she found that in an academic research design, she had to make the questions more specific and complexify both the concept of background knowledge and the specificity of texts in the process. Meanwhile, test development and use proceeded. Clapham's results were used eventually in combination with practicality concerns to make further test development decisions. This combination of desirable conceptual properties and practical constraints of implementation was a recurrent theme in the development and validation reports. This is not intended to reduce the value of the IELTS development and validation reports. They formed a necessary basis for the operational use of the test.

7.5 Post-publication reports on IELTS development, validation, and use

The research on the operational IELTS that has been published so far concerns test development, validation, and test use. The topics include ongoing test development, quality monitoring and the specific nature of the construct being assessed, concurrent and predictive validity, and acceptability and spread of the IELTS test.

7.5.1 Operational test development

In April 1995, some changes were introduced into the IELTS test. There had been three academic reading and writing modules which, within the academic speciality, combined the reading and writing tests. After the change, there was one reading test and one writing test for all academic candidates and the reading and writing modules were no longer linked. The second main change was that whereas the General Training module had only reported scores up to IELTS band 6, they were now to be reported on the same 9-point scale as the academic version of the test (UCLES 1996:7).

Charge and Taylor (1997) described the changes in IELTS and reported the rationale for them. Although the thematic link between reading and writing was desirable in some senses, it also made assessment more difficult because some writers made extensive use of the link while others did not. Furthermore, while some writers showed their ability to apply the information from the reading task into their writing performance, others made such extensive use of the reading text that their own writing was not very evident in their performance. The removal of the link also made test administration easier while also making it more practical to produce comparable versions of the writing section since the thematic link no longer constrained the selection of writing tasks.

Charge and Taylor (1997) also described the administrative changes that the revision in April 1995 included. Centres were allowed to schedule the speaking section for some candidates up to two days after the other test sections to help administrative pressures at large test centres. The schedule of the listening section was amended to allow candidates time to transfer their replies to the optical marking sheet. Furthermore, a candidate information sheet was added to the IELTS administration package, a computer program was released to help test centres administer and organise test sessions and record performances, more test versions were made available, and a service was made available whereby candidates could formally query their results. The data gathered on candidate background

were to be used in quality monitoring and in validation research through more detailed analyses of differential item functioning for subgroups of candidates.

7.5.2 Test monitoring and maintenance

Research related to test monitoring and maintenance focuses on the procedures which are routinely implemented in operational testing. Its main aim is quality monitoring. In the IELTS literature, there are three research reports that concern this area.

Coleman and Heap (1998) focused on possible misinterpretation of rubrics in IELTS tests of listening and reading. The study arose in response to criticisms from some language instructors, who had found some of their students' IELTS scores baffling, and from some IELTS invigilators, who had raised concerns that the rubrics might be difficult to comprehend. These doubts posed a hypothesis that some of the IELTS candidates may receive low scores because they do not understand the test instructions and not because of their language ability (Coleman and Heap 1998:39). The researchers set out to investigate this hypothesis.

The researchers analysed candidate responses to reading and listening items from an operational candidate performance database, and interviewed a group of 13 students to find exactly what they understood the rubrics to mean. Their overall finding was that relatively few students had any difficulties understanding the rubrics (1998:70). They only recommended two clarifications to existing rubrics and encouraged the testing board to monitor even more carefully that the standardised rubrics are always used on all test forms.

However, Coleman and Heap (1998) did find other areas where the test development and administration procedures could be improved. The wording of the actual *test questions* appeared to have caused some confusion, especially questions which contained a negative or a double negative. The researchers recommended that such questions should not be used. Furthermore, frequent variation in question type appeared to be confusing for some test takers, so they recommended against this. Lastly, they found that the markers did not always follow the marking key very strictly, and pointed out that a mechanism for detecting which markers marked which papers would help monitor this and might also increase marker accountability. (Coleman and Heap 1998: 70-71.)

A study into the way the IELTS speaking test operates was conducted by Brown and Hill (1998). The researchers pick up one of Ingram and Wylie's (1993:229) proposals for further research to investigate

the relationship between interviewer style and candidate performance in the IELTS oral interview. Some 32 candidates and six interviewers took part in the study, each of the candidates going through two IELTS interviews with different interviewers. The results showed that although the interviewers followed the same scripts, they used different strategies, and sometimes this resulted in the award of different grades for individual candidates. The easier interviewers tended to shift topics more frequently and asked questions that requested simple factual information or description (Brown and Hill 1998:13). More difficult interviewers asked more challenging questions, interrupted the candidate, and sometimes disagreed with them (pp. 10-18). Brown and Hill (1998:18) suggested that the more structured an interview is as a question-answer routine, the easier it appeared to be. Furthermore, Brown and Hill (1998:3) suggested that it was possible that actual differences in interviewer behaviour were quite big but that this might be masked in interview grades because there is evidence from other studies that raters may compensate for interviewer 'difficulty' in the grades that they give.

The upshot of their results for the IELTS board, Brown and Hill (1998:18-19) suggest, is that the board must decide whether they want the 'easy' or 'difficult' style of interaction to be tested in the IELTS interview and then to train interviewers to make sure that similar interactional challenges to candidates are presented by all IELTS interviewers. They stated that the more 'difficult' style with interruptions and disagreements may be closer to 'natural' conversational behaviour than the 'easy', supportive style. They proposed awareness-raising, monitoring, and self-monitoring as strategies for ensuring that interviewers behave in the desired manner and help the same skills be tested with all IELTS participants.

Merrylees and McDowell (1999) continued the research strand on the speaking module. They conducted a survey of IELTS examiners' attitudes towards the test and made a preliminary analysis of 20 transcribed interviews to investigate how examiner discourse affects the quantity and quality of candidate discourse. The survey focused on the examiners' attitudes towards the interview format and their attitudes towards, and use of, the IELTS band descriptors. The analysis of transcripts was limited and only comprised analyses of length of turn in terms of numbers of words and numbers of minutes and seconds.

The results indicated that the majority of the examiners were comfortable with the IELTS structured interview, but that a minority group of examiners would rather see the current format changed (Merrylees and McDowell 1999:10, 26). They found the assessment scale in need of some

revision, especially clarifications to scale descriptors at levels 5 and 6, and possibly an addition of analytic or profile scales in addition to the overall speaking scale. They proposed that examiner training include regular reminders of how the interview should be conducted to make sure that all examiners follow comparable procedures, that scale descriptors at levels 5 and 6 be clarified, and that some form of relatively close examiner monitoring be implemented.

The analysis of transcribed interviews indicated that Phase 3, a role play where the candidates ask questions of the examiner, resulted in the examiner talking much more than the candidate. Merrylees and McDowell (1999) appeared to construe this as a negative finding. Phase 4, a topic-focused interview, seemed to elicit the largest amount of language from the candidates in relation to examiner talk, which the researchers found appropriate. However, a complicating factor was that the amount of examiner talk in this section was not necessarily related to the candidate's level. Several examiners of level 6 candidates talked a great deal during this phase while examiners of level 5 candidates talked less and allowed the candidate more time to talk. The researchers interpreted this to mean that the examiners were unnecessarily scaffolding the level 6 candidates' performance, while the level 5 candidates' examiners were more challenging when demanding candidate talk. This is not immediately evident from the numerical analysis of words per turn, but the researchers did of course have access to the content of the transcripts, not just the numbers reported in the paper, so the interpretation may be accurate. Be that as it may, this quantitative difference in examiner behaviour led the researchers to question the reliability of IELTS examining procedures, as it may indicate that the test does not challenge all candidates in an equal way. Another feature that varied between different examiners was the amount of time spent on the different phases of the interview. Some spent more time on the initial familiarisation phase, others on Phase 4, which was discussion on a topic. Merrylees and McDowell (1998:34) recommended that this lack of standardisation should be addressed at examiner training sessions.

The type of research summarised above is useful for monitoring the need for change. This constitutes part of quality assurance, which Bachman and Palmer (1996) have termed "evaluation of test usefulness". The studies are also integrally related to test validation, investigating as they do the actual testing process. The data gives evidence of how the operational procedures implement the intended construct, and the results also show what kind of variability there is in the operational procedures. Implications can then be

drawn about whether the test is testing appropriate constructs in an acceptable way, and whether the variability is a threat to test validity.

All the above studies on IELTS led to a recommendation to tighten the examination procedures. Alderson et al. (1995:227-228) suggested that examination boards had two main avenues of action open to them: they could keep the examination stable for a number of years and then implement a large scale revision, or they could implement small changes as and when the need arose, so that the need for major changes would be reduced. Whichever the board's decision, the aim remains to keep up quality standards and make the examination useful for its users. Alderson et al. (1995) did not consider the measurement equivalence of slightly altered forms, although this would be one of the quality concerns addressed. However, in the case of the IELTS decisions discussed above, if the aim of the revision were to improve a measurement quality that had been found deficient, it would probably count as a desirable development.

7.5.3 Aspects of IELTS validity

The two studies on operational IELTS which are explicitly labelled validation studies investigate predictive validity, or, in the terminology used in Chapter 3, predictive power. Other studies on IELTS which can be seen as part of the modern, broad concept of validity focus on scores and score comparability, impact, authenticity of the test tasks, and acceptability of the test to its users.

7.5.3.1 Predictive validity

Cotton and Conrow (1998) conducted a classic, situated predictive validity study of IELTS at the University of Tasmania. They followed up 33 international students to see how well the IELTS total scores and subscores predicted the students' academic success and any language problems that they might have experienced during their first year of study. Furthermore, they surveyed academic staff to see how well IELTS scores predicted staff ratings of student performance. The data consisted of the 33 students' IELTS scores, their responses to a questionnaire, interviews with 23 students, and staff surveys sent to academic staff, international student advisors, and tutors on English support. In addition to prediction of success, Cotton and Conrow endeavoured to list other key variables which appeared to have the most effect on academic success besides language ability.

Cotton and Conrow's (1998) results indicated that the relationship between IELTS scores and various indicators of academic success as well

as experiences of language problems was quite weak. The correlation coefficients with academic results were sometimes negative or close to zero, ranging from -.58 for listening with semester 1 results to .42 for reading with the full year's academic results (Cotton and Conrow 1998:93-94). However, the correlations were based on 17 and 26 students' performances respectively, so absolute values should not be trusted. Nevertheless, they indicate the direction of the results. The qualitative data gathered through questionnaires and interviews suggested that academic achievement is influenced by a multitude of factors in addition to language ability, including the amount of English language assistance received, motivation, cultural adjustment, and welfare difficulties experienced (Cotton and Conrow 1998:110).

Cotton and Conrow (1998:98) admitted that their study suffered from the problem of the truncated sample so common to studies of predictive validity: the students who were not admitted to the University of Tasmania, perhaps because of their IELTS score, were not included in the study population. This may have been one of the explanations for the low correlations. Other explanations that the researchers listed included cultural responses to self-evaluation, varied perceptions of what constitutes academic success, culture shock, and differences in academic expectations between the home culture and the Tasmanian academic environment (p. 97). They recommended that predictive validity studies should be conducted with larger, more homogeneous samples and that intervening variables should be studied with carefully designed instruments of observation. To the IELTS board, the researchers suggested that information about the examination should be disseminated to academic staff so that they would learn what information IELTS scores could offer them.

Hill, Storch and Lynch (1999) compared the effectiveness of IELTS and TOEFL as predictors of academic success at the University of Melbourne. The researchers' data consisted of 55 students' first semester course grades and their overall scores on either IELTS or TOEFL, questionnaire responses from 66 students, and interviews with 22 volunteer students. The research questions focused on prediction of success and the role of other factors, such as English language support, in facilitating academic success (p. 54).

As common in predictive validity studies, Hill, Storch and Lynch's (1999:55) results indicated that the predictive power of IELTS overall score for first year grade point averages at university was moderate with a correlation coefficient of .54 and that the predictive power of the TOEFL overall score was weak with a correlation of .29. They suggested that part

of the explanation might be that the students' Test of Written English scores were not included in the comparison as they had not been available from the database they used (Hill et al. 1999:61). Regarding language support, the researchers found that those seeking it had lower test scores and lower grade point averages for the first year – in other words, those who most needed the language support also sought it (1999:62). The questionnaire responses indicated that the reasons for seeking language support and the relationship between English language ability and academic performance were complex. Students reported different reasons for contacting language support, different expectations of effectiveness, and different self-assessments, among other things. Hill et al. (1999:62) concluded that while the complexity of factors influencing course success explains why the degrees of prediction are not higher, examinations are nevertheless helpful in identifying those students who are most in need of language support.

7.5.3.2 Scores and score comparability

Celestine and Cheah (1999) conducted a study on the effect of background disciplines on IELTS scores. Curiously, the authors do not refer to Clapham's studies on the topic, though this may be explained by the slightly different and tightly controlled focus of their study. The researchers investigated the effects of previous education, which in this study was either the Science or the Arts stream of Malay secondary school, on students' scores on the IELTS test. The students were matched on their grades from the secondary school leaving examination in English, and the researchers investigated the possible differences in the students' IELTS scores.

Celestine and Cheah (1999:44-46) found that, overall, a background in either the Arts or the Science stream in Malay secondary school did not make a difference in the IELTS scores. However, when they investigated the scores in finer detail, they found that there were some statistically significant differences in the way intermediate or weak English learners did on the IELTS. The students with a Science stream background did better. The researchers explained this by two factors: firstly, students in the Science stream tend to be academically more able in general and, secondly, the kinds of learning styles supported by the science stream might explain the students' higher scores. By deduction from the researchers' explanations of strategies supported by the Malay educational system, they seemed to suggest that ability to solve problems, hypothetico-deductive reasoning, ability to apply ideas practically, ability to single out best or correct answers, and a systematic and scientific approach are helpful strategies for performing on IELTS. No validation of this proposal was attempted.

Celestine and Cheah (1999) also looked into the effectiveness of IELTS practice courses, and found that the kinds of short courses that the Malay students took had no statistically discernible effect on their scores on any ability level. This makes an interesting contrast to Brown (1998; see below), who found that ten-week intensive IELTS preparation courses *were* effective in raising students' scores on the writing section of the IELTS test.

Mok, Parr, Lee, and Wylie (1998) investigated the comparability of the IELTS scale with that of the ACCESS test, which was used by the Australian Department of Immigration and Multicultural Affairs to assess intending migrants' language skills. The motivation for the study was that IELTS was an alternative examination for immigration purposes and it was necessary to see to it that the same standards could be kept by both examinations.

The study was conducted by means of a linking scale, that of the Australian Second Language Proficiency Rating (ASLPR). This method was selected because, given the costs of the two examinations and the situation in which their immigrant takers are, it was impossible to set up a rigorous research design where a sample of the target population would take both tests. As the researchers say in their discussion (Mok et al. 1998:163), the sample was opportunistic. Although the sample sizes within each of the examinations involved in the design were reasonable with a range from 355 to 759, the number of the candidates who had taken more than one of the tests was very low: 32 altogether. Moreover, the researchers pointed out that only regarding the ACCESS and the IELTS General Training modules was there adequate spread of centres to allow generalisation of the scale to the whole population of takers of the respective tests without influence from particular assessors at particular centres.

When they attempted to compare the ACCESS and IELTS scales, Mok et al. (1998:161-163) found that the subskill-specific scales within each test were different from the overall scale, and this led them to establish correspondences between the tests skill by skill. They estimated the overall correspondence and concluded that ACCESS level 4 seemed to correspond to IELTS range 5.5-7, but they considered the macro-skill-specific scales and correspondences more accurate. This was because the tests differed in the relative difficulty of the sections. In the discussion, Mok et al. (1988:165) thus warned against simplistic equating between tests according to overall levels and recommended macro-skill-specific equating as a more accountable way of establishing correspondences.

Mok et al. (1998) admitted that their research design was problematic and they clearly reported on the many regions where the scales that they

investigated did not fit the statistical model they used. They discounted criticisms based on the small number of participants who took two of the tests, but I would suggest further caution in this area when the results are interpreted. The researchers argued, however, that even with its flaws the merit of their study was that it was empirical and based on actual test data (Mok et al. 1998:163-164). It thus complemented the most common approach to the comparison of examinations and scales, which is perception-based. While this is true, the implementation of the study leaves much to be desired and calls for a more careful research design. However, the researchers' conclusion raised an important validity question for Australian immigration authorities about fairness of score use, namely the empirical comparability of cut scores used for different tests.

7.5.3.3 Impact and authenticity

Brown (1998) conducted an evaluation of two language courses at the Hawthorn English Language Centre in Australia. His study was relevant for IELTS because one of the courses was an IELTS preparation course; the other was a generic course on academic writing. The criterion measure used was the IELTS writing section. Brown found that IELTS preparation was efficient: over a ten-week intensive study program, students in the IELTS course improved their IELTS writing performance much more than students on the generic EAP course. The groups were very small, N = 9 in the IELTS group and N = 5 in the general EAP group, and moreover, the ability level of the IELTS group was lower at the start of the program. The Mann-Whitney U test to examine the significance of score differences after adjustment of ability was significant at p = .0441, which formed evidence against the null hypothesis of no difference between the groups.

The aspect of Brown's results which is particularly interesting to the validation of IELTS is his explanation of what it was that made the IELTS students gain more over a ten-week period. According to Brown (1998:36), the students gained because their skills in writing and planning their writing developed, their teaching focused on the IELTS assessment criterion of writing task completion, and they were trained in strategies for writing under examination conditions, especially timed writing. Brown cautioned that the sample was small and therefore the results may not be generalizable. However, he also stated that to the extent that a difference did exist, the gains were made in IELTS performance. The larger gains of the IELTS group on this measure did not necessarily mean that they became better academic writers than the other group over the 10-week intensive course. This could only be assessed by using some operationalization of

"successful participation in Australian tertiary education" as a criterion measure. The alternative explanation was that the IELTS writing scores were to some extent susceptible to coaching effects which might or might not benefit academic writing more generally.

Moore and Morton (1999) investigated the authenticity of Task 2 in the writing section. They compared 155 writing assignments from various university departments against a corpus of 20 Task 2 writing items. They analysed the tasks in terms of genre/text type, rhetorical function, the object or topic that the task focused on, and the source of ideas that the writers were instructed to use, whether own prior knowledge or primary or secondary sources. Additionally, Moore and Morton conducted a staff survey in which they asked 20 staff members who had provided material for the study about the nature of their writing assignments, and asked them to compare their own writing assignments against the IELTS Task 2 items.

Moore and Morton (1999) found that the IELTS tasks corresponded to university tasks in terms of genre in that the most frequent task in both contexts was the essay. However, they also found clear differences; for instance, IELTS items asked the writers to use their own ideas and prior experience, while university tasks called for the use of a variety of research techniques. Furthermore, the range of rhetorical functions in IELTS was restricted, and the topics were often concrete, whereas university tasks focused on abstract entities. The staff survey provided supportive evidence for these findings.

The researchers considered their results in the light of the demands of test-based writing, and recognised the fact that full authenticity was not possible. However, they presented a range of recommendations through which the authenticity of IELTS Writing Task 2 could be improved. These included the re-introduction of a link between the writing task and one or more of the reading tasks, which had also been called for on similar grounds by Wallace (1997). Furthermore, Moore and Morton proposed more frequent use of prompts requiring summarisation, comparison, explanation or recommendation rather than an examinee's opinions on the desirability of a social practice, and an increased emphasis on abstract, idea-based discourse through the use of prompts which focus on other people's ideas. Prompts should begin with wordings such as 'many psychologists argue' or 'some educationalists believe'; in this way, the examinees would be obliged to write about other people's arguments using more complex language than they would need for the expression of their own views (Moore and Morton 1999:100-102). Moore and Morton also suggested that such modifications

to IELTS writing might enhance the washback effect of the examination to the content of teaching in preparatory courses (p. 103).

7.5.3.4 Acceptability and test use

McDowell and Merrylees (1998) investigated the degree to which Australian institutions of higher education used IELTS as their language qualification and surveyed their perceptions of the IELTS test. The results indicated that IELTS was the most commonly used language test among Australian institutions and also the preferred test (1998:120, 136, 138). Furthermore, the survey revealed that a range of in-house tests were used for the assessment of students' proficiency. The institutions reported that they used fixed language requirements in admission, but that a range of decision making bodies were consulted for the minimum language requirements. The sources consulted included language professionals, educational testing literature, and other educational institutions. McDowell and Merrylees (1998:139) appeared to treat their research partly as a publicity exercise and listed as one of their conclusions that the survey as such has raised the profile of IELTS among the higher education community.

7.6 Case summary

I will summarise the report on the IELTS case by answering the questions in the case study protocol. They focused on test development, validation, theoretical construct definition, and values that guide test development and validation.

The developers of IELTS identified six steps in the initial development of the test. These were evaluation of the starting point, production of draft specifications and tasks, gathering of stakeholder reactions to the drafts produced, preparation of final specifications and production of trial tests, pilot administration and analysis of data for development and validation, and production of final forms of the test and related information and training materials. The process relied on existing literature, stakeholder advice, trial implementation and framework revision, and collegial collaboration.

The process of IELTS development illustrated the iterative nature of the activities. For instance, when draft tasks were produced, small scale trials were arranged and data analysed to examine how good the test was. This was assessed through comparability of scores, tasks, and constructs that appeared to be measured. Improvements were made in tasks, test specifications, and construct definitions as needed. Changes were made as

late as after the main trials when it was found in a correlation and factor analysis that the grammar test did not add a new component to the test.

The initial development of IELTS appeared to be guided by the desire to measure the right thing and the desire to serve stakeholder needs and wishes. The needs were investigated and detailed descriptions of measurement intent as well as draft tasks were developed to clarify the intent and verify the acceptability of the forthcoming test. Comparability between test forms was investigated through trials, and when it was found that it did not meet the requirements of the project, the specifications were revised to develop better comparability. Measurement economy was heeded when the main trials revealed that it was possible to delete a section of the test without compromising its measurement quality. One practical result of the detailed description of measurement intent was that the information could be used in the creation of information material for the test. This material was used in the introduction to the case when the nature of the test was described.

Development reports related to the operational stage of IELTS focused on quality maintenance and empirical validation. Whereas the pre-publication reports and articles had discussed the development from the developers' point of view, the post-publication reports did not seem to be written by participants. The recommendations that the researchers made indicated that the decisions would not be made by them. This is partly a decision of publication policy, but it is probably also the logical result of test publication. After it, the aim is to produce comparable versions of the same test and improvements are desirable when the properties of the examination are found to be so variable that they threaten comparability. While the measurement comparability of test forms was probably monitored, information on eg. the reliability of the speaking and writing modules was not published. Compared with the framework of test development presented in Figure 3 in Chapter 5 above, the IELTS case possibly implemented all the components of operational test development, but short of Charge and Taylor's (1997) summary of decisions, they did not report on it from the developers' perspective.

Validation during the initial development of IELTS was clearly focused on the development of the content and format of the test with the help and simultaneous clarification of the construct definition. By using their own experience in the development cycle, the test developers sought an inductive approach to the definition of the skill assessed. They argued that the records of their test development rationales and decisions, which included construct descriptions, constituted validation evidence. This was provided in the form of experiential knowledge and reports of measurement intent and

stakeholder views. Such recording of test development procedures is recommended in the current *Standards for educational and psychological measurement* (AERA 1999:43).

The test developers' approach to validation during the initial stage seemed to combine the questions of "What do we want to measure?", "What can we measure?" and "How should we measure it?". The openness was constrained by the test development brief, which required the reporting of scores for four skills and the use of an existing reporting scale. The work focused specifically on construct description and the development of a test of "the right thing". In trials, the reliability of the reading and listening sections was investigated, but the measurement properties of the test were not emphasized as much as the conceptual validity questions.

During the operational stage of the test, the validity questions became more independent of test development and more concerned with score explanation and score use. Predictive validity, score comparability, score use, and impact became relevant issues, as they should in an operational test. The data used in these studies was scores, background information on examinees, and external indicators for examinee ability. The results indicated the degree of usefulness of the scores in the context investigated and usually also included advice for how the research design of similar studies could be improved in the future. Questions that combined test development and validation were also asked, especially by Moore and Morton (1999) in their study of the authenticity of the writing task. Similar implications could have been drawn from Celestine and Cheah's (1999) study if the questions of the constructs that they proposed to explain the scores had been investigated.

In relation to the areas of test-related validation identified in Figure 3 in Chapter 5, the validation discussions related to IELTS addressed the majority of them. During initial development, the evolving system was constantly evaluated against aims and the state of theory, and plausible rival hypotheses were identified if not formally experimentally tested. During operational development, the appropriacy of test use was addressed in a number of studies, evidence for score meaning was gathered and analysed, and at least two contextualised validity studies were conducted. The sample sizes in the studies were small; this was recognised by the authors, however. Sometimes the researchers proposed revisions to the test, but the diction indicated that they would not be making the development decisions, they simply made recommendations. One small-scale study (Brown 1998) touched on the impact of IELTS on teaching, but further studies more clearly focused on the possible and actual impact of IELTS should be conducted. Plans of future validation research were not published during the

initial stages of test development although discussions may well have been conducted among the developers and at the examination board. Furthermore, areas which were not taken up in the publications related to IELTS validation at all included the examination of values that guided test development. This would require a critical approach, and such studies can perhaps not be expected to be written by the people who are in charge of the initial development of an examination developers or published by examination boards who are responsible for its financial viability. However, a forum for such critical appraisals should perhaps be identified, and one possibility might be the language testing research community.

The construct assessed in the IELTS test was defined briefly as degree of readiness to study or train in the medium of English (UCLES 1999: inside cover). The *IELTS Handbook* also gave some detailed characterisations of what was assessed in each test section and even in each task in the writing section. These were summarised at the beginning of the present case report. The definitions characterise the abilities expected of the examinees, but they also concern the topics and text types that the test is likely to include and they describe in general terms the operations and activities that the examinees will perform. This definition corresponds to Chapelle's (1998) category of interactionalist construct definitions. Both the individual's ability and the language use context are described, and in the interest of characterising the ability in a relevant and useful manner, several aspects are identified within each of these. The definition was developed in the course of initial test development. It is operationalized in test development through test specifications. Its operationalization in validation studies is not equally clear for someone who is not a member of the testing board. There are traces of a validation program in the IELTS research reports and studies, but no comprehensive presentation of it. Such a program would have to face difficult questions about the interface between the theoretical and the psychometric construct definition. The study that was published on the main trials of the IELTS test indicated that the test could be considered to measure a single underlying dimension. The way in which the complex construct definition sits with this measurement result is an intriguing question that has not yet been addressed in the publications related to IELTS. As regards measurement quality, there are very few publications on this aspect of the test. It is therefore difficult to judge how this aspect of construct definition is operationalized in the development and validation of IELTS.

Judging by the test format, the developers of IELTS supported a communicative or, more precisely, an interactionalist view of language with

an emphasis on actual language use. That is, it was considered important to assess writing through productive writing tasks and speaking through a structured interview. This meant compromises in measurement quality, because double marking was not a standard requirement and variation in test administration was detected even though the interlocutors supposedly followed the same script. This is a value decision on the part of the testing board, who must have considered the assessment of productive skills so important that the compromise of measurement quality was possible. The theoretical argument to support such a decision would likely be construct representativeness. Whether or not this is true, the case data shows that construct representativeness was not always the main criterion followed when test development decisions were made. When the developers decided not to test reading and writing in an integrated manner, the main reason was reportedly the test development brief, which required that scores must be provided for the four skills.

Judging by the assessment procedures followed, quality in assessment was ensured by the detailed assessment criteria used and the expertise of the judges employed; many of the items were open-ended and required human judgement, and double marking was not standard. The alternative might have been the use of psychometric criteria through selected response tasks and the employment of at least two raters, but this was not done in IELTS. The activities of the assessors were reportedly monitored by spot checks, but the results were not published. A study on rubric interpretation (Coleman and Heap 1998) produced the incidental result that the markers of open-ended items were not completely consistent, which should send a warning signal to the test developers about reliability of marking. Reliability studies of the writing or speaking sections of IELTS were not published.

The studies published on the development and validation of IELTS indicated that professional expertise was the main justification for the development decisions made. During initial test development, the professional expertise undoubtedly came from the past experience of the participants but they also used their experience of the actual test development process when they recorded the test development process, which contributed to validity evidence for the test. Development decisions were also made on the basis of combined utility and psychometric criteria when the grammar test was excluded, and on the basis of fairness and practicality considerations when the subject specialisations of the reading and writing modules were removed. Overall, however, the measurement quality of the test was not emphasized according to the studies published.

8 EXTENDED THEORETICAL AND PSYCHOMETRIC DEFINITION OF CONSTRUCT: TOEFL 2000

8.1 Introduction to the TOEFL 2000 case

TOEFL 2000 is an ongoing, broad research programme at the Educational Testing Service to revise the TOEFL test for the twenty-first century. According to the project's mission statement, it will "deliver to the TOEFL program an English language assessment of communicative language ability that is valid, cost effective, easily accessible, and efficiently delivered" (ETS 2000c). The project framework is intended "to improve the measurement of second-language competence" so that "rather than merely assign numerical values or positions to examinees based on their responses to a set of tasks, the goal is to give meaning and interpretability to the numbers" (Jamieson et al. 2000:7). TOEFL 2000 is governed by the TOEFL Policy Council at ETS. Its members include ETS staff and external consultants (ETS 2000c).

Not being an insider of the TOEFL test nor having a history of administrative records to analyse, I will summarise the history of the project on the basis of the accounts that have been published in publicity documents related to the project. The actual development history may have been much more complex, but the internal politics of the TOEFL examination body is not the topic of this thesis. According to the TOEFL 2000 website (ETS 2000c), the TOEFL 2000 project was founded in 1993 in response to wishes from score users, applied linguists, language testers, and second language teachers that the traditional paper-based TOEFL test should be replaced with a new test which would be "more reflective of communicative competence models", include "more constructed-response tasks and direct measures of writing and speaking", contain "tasks that integrate the language modalities tested", and "provide more information … about the ability of international students to use English in an academic environment" (ETS 1998:6). To tackle these challenges in earnest, the project staff undertook research in three broad areas: the needs of test users, technology for test design and test delivery, and test constructs (ETS 1998:6, Jamieson et al. 2000:4). They conducted surveys and systematic reviews of literature, and worked on a model of communicative competence for the new test. The results with attempts to implement the model showed that the conceptual work on TOEFL 2000 needed a longer development time than was originally envisioned. The decision was made to move "the

current TOEFL with some important design enhancements to computer-based testing in 1998 while continuing to pursue the original vision of TOEFL 2000 within the Research Division of ETS" (ETS 1998:6). The continuing conceptual work is the focus of the present case.

8.1.1 Boundaries of the TOEFL 2000 case

In terms of time, the TOEFL 2000 case encompasses published project work from the start of the project to the present day. In terms of stages of test development, the case covers the first steps of initial test development, where the grounding of the test in its need and purpose is clarified and the test construct and format are defined. With reference to Figure 3 in Chapter 5 above, the TOEFL 2000 case covers early developments in the top half of initial test development. Given that the computer-based TOEFL test (CBT) is currently available and the TOEFL 2000 website states that the new test will not replace it but the CBT is a first step in the development of TOEFL 2000 (ETS 2000c, *frequently asked questions*), it could be argued that the TOEFL 2000 test has been published and it is in operation at the moment. However, because the conceptual structure for the test is still under development and it is not clear what sections the new test will contain, I will conservatively consider the development to be at an early stage where many realisations of the test format are still possible.

I will base the case analysis on published research reports in the TOEFL Monograph Series. To guide the selection of reports for analysis, I will use Jamieson et al.'s (2000:51-52) identification of the report foci into test constructs, psychometric models and procedures, trends in international student enrolments, and technology applications. Although the technological developments of the TOEFL 2000 project are interesting, they are not central to my research. Accordingly, I will discuss them only to the extent that they are taken up in the reports on conceptual development that have been published. For reasons not explained in any of the publicity material, the publication of the research report on trends in international student wenrolments is on hold. I therefore will limit the analysis to test constructs and psychometric models.

8.1.2 Format of the TOEFL 2000 test

The exact format of the TOEFL 2000 test has not yet been established. Given the *TOEFL 2000 website* statement that "the TOEFL 2000 project now continues with efforts designed to establish a solid foundation for the next generation of computer-based TOEFL tests," the medium for the test is the computer. Furthermore, the list of project goals indicates that the

skills of reading, writing, listening, and speaking are likely to be assessed in it both independently and interdependently.

8.1.3 Developers of the TOEFL 2000 test

As stated above, the administrative control of the TOEFL 2000 project is with the Research Division of ETS rather than, for instance, the Test Development department. The project website (ETS 2000c) shows a complex management structure, where the management level includes a project management team and a number of advisory committees to govern the work of eight working groups. The committees comprise the TOEFL Advisory Committee, an ETS Oversight Committee, a Research and Development Oversight Committee, and a Research and Development Advisory Committee. Of the working groups, four are skill-specific, one each for reading, writing, listening and speaking. The teams are composed of ETS test developers, researchers, and external consultants, who represent areas such as first and second language teaching, instructional design, technology applications, and language testing research (Jamieson et al. 2000:39). The other four teams are for prototyping, psychometrics, technological capabilities, and marketing (ETS 2000c). The existence of these groups shows an awareness that specialist knowledge is needed to implement and market innovations however good they should be. However, it also means that the innovativeness of the research and development groups is constrained by practical limitations of time and money.

8.1.4 Test development brief: conditions and constraints

The *TOEFL 2000 Framework* (Jamieson 2000:1-2) defines the scope of the TOEFL 2000 project through its constituencies, constraints, and project framework. Constituencies refers to the groups of people who are affected by the test; the IELTS developers called them stakeholders. This encompasses admissions offices and other administrative score users, teachers of admitted students, and test takers. The different needs of these groups should be served by the meanings available from the scores of the new test. The constraints are practical, including availability of computers if the computer is the sole medium of delivery, and the need to provide a secure test on demand to very large numbers of people all over the world. The framework defines the test's purpose, its task characteristics, and the conceptual framework that underlies the task characteristics (Jamieson et al. 2000:7). The nature of this framework will be discussed in the present case. The analysis will clarify what it means that a test's construct definition is extended both from a theoretical and from a psychometric perspective.

According to the *TOEFL 2000 website* (ETS 2000c), the goals of the project are to provide:

1. measurement of broader constructs than the current TOEFL test,
2. assessment of the four modalities of reading, writing, listening, and speaking,
3. more constructed-response tasks and direct measures of writing and speaking,
4. integrated tasks and modeling of their relationships,
5. more precise definition and scaling of the four skills,
6. more extensive information to test users (including examinees) about test performance (ie., performance descriptors),
7. positive impact of the test on instruction, and
8. a fiscally sound testing program that is market-sensitive and accessible.

http://www.toefl.org/toefl2000/

The first seven goals define the conceptual improvements over the existing TOEFL test, of which some – such as increased number of constructed response tasks – are more concrete and easily evidenced while others – eg. positive impact on instruction – are less tangible and in need of concrete means to work towards the goal. The last defines the practical goals of the project in generic terms, except for one important consideration, that of development time. Whether or not there are limits on this, these have not been published.

8.2 Nature and focus of studies discussed in the TOEFL 2000 case

In this section, I will briefly characterise the nature and focus of the studies that will be analysed in the present case. These have all been published by the ETS, most of them in the Monograph Series. This means that the perspective is the test developers' as mediated by the examination board-cum-publisher. Other reports on the development have to my knowledge not been published, possibly because the development is in such early stages that external trials eg. with prototypes have not been possible, but possibly also because confidentiality may bind those who know about the development not to publish about it. I will take up the issue of perspectives on test development in Chapter 9.

In the Foreword to the reports in the TOEFL Monograph series, the TOEFL Program Office states that the reports published in this series "were commissioned from experts within the fields of measurement and language teaching and testing" and that they "have helped to inform TOEFL program development efforts with respect to test construct, test user needs, and test delivery" (eg. Jamieson et al. 2000:i). However, the Program Office makes it clear that the views in the papers do not necessarily reflect the views or intentions of the TOEFL program. The analysis here concerns the contents of the reports, and this statement is a reminder that the decision-making in

the project is not part of the material that I analyse. This is unfortunate, but as an external analyst I must rely on published materials.

It is clear from the contents of some reports that they have been written by people actively involved in the development of the new test. This is especially so for two papers, namely Chapelle et al.'s (1997) presentation of the model of communicative language ability developed by the TOEFL Committee of Examiners (COE), and Jamieson et al.'s (2000) *TOEFL 2000 Framework*, which reports on project history, describes the development rationale, provides a conceptual framework for the project, and reports on the stage of progress in working through the agenda. These are among the key reports to be analysed in the present case.

A table of the reports published on the conceptual development of TOEFL 2000 can be found in Appendix 5. It identifies the reports by number, author and date, and additionally characterises the focus of each report, the materials and methods, the main conclusions, and the area of test development and validation that each report concerns.

Four groups of studies can be distinguished. The first focuses on the domain or needs analysis of communication in academic life. The studies judge the relevance and adequacy of the available literature to form a basis for TOEFL 2000. The second group focuses on reviews of theoretical literature on constructs that are relevant for the new test. These reports include the Chapelle et al. (1997) paper mentioned above. This will be discussed in depth, whereas the contents of the skill-specific construct papers will be discussed more briefly with specific focus on their contribution to the construct definition and possible realisations in TOEFL 2000. The third group encompasses the Jamieson et al. (2000) *TOEFL 2000 Framework* and four related skills-based framework documents. Since these follow a standard outline and only two of them have been officially published, only these two will be discussed in the case report. The framework documents define the construct assessed in more concrete and more readily implementable terms than the theoretical literature reviews. The fourth group concerns the measurement implications of TOEFL 2000. Since no extensive trials have been conducted yet, the measurement papers are a review of issues from measurement literature and sample analyses with data that could be representative of the forthcoming test. In addition to these four groups, I will allocate a separate subheading for TOEFL 2000 validation work in the case report.

All the reports that have been published so far on the non-technological aspects of TOEFL 2000 are based on background theory. Together, they form an arc from theoretical considerations to frameworks

nearing readiness for operationalization. The topics addressed in the reports are those that the TOEFL 2000 project considers necessary for construct definition: communicative competence, academic context, possible impact, and measurement aspects. The constructs are defined through words that name the theoretical concepts and important variables. In addition, the frameworks outline possibilities for empirical connections between theoretical and psychometric construct definitions.

8.3 Initial development of TOEFL 2000

8.3.1 Analysis of communicative needs in academic contexts

Two early research reports in the TOEFL Monograph series concentrated on the characteristics of the domain of language proficiency that TOEFL 2000 is intended to assess, namely English for Academic Purposes in the United States and Canada. Waters (1996) analysed needs analysis studies that had been conducted in North America and Britain on second language speakers. Ginther and Grant (1996) analysed studies on the language needs of native English speaking students at undergraduate and postgraduate levels. The commissioning of the studies was based on the rationale that if it were possible to categorise academic language use tasks in detail, this categorisation could guide test development and form a basis for validation (Ginther and Grant 1996:1). To provide such a framework, the report writers should identify tasks, analyse their parts in terms of student behaviours, establish the frequency of different behaviours, and study instructors' evaluations of the tasks.

Waters (1996) reviewed existing research into needs in English for Academic Purposes. He identified several dimensions in the existing literature, including division into skills versus investigation of language use situations and activities where skills are integrated, research on native and non-native speaker needs, and studies concerning perceived needs versus observations of actual academic situations and analyses of texts. He noted that sometimes analyses had been conducted on faculty perceptions, and at other times different categories of students had served as informants. Some studies concentrated on undergraduate needs and others on postgraduate needs. Waters's results indicated there was not a comprehensive research literature on the needs for Academic English at North American universities which could have informed test development. Waters was aware of the problematic nature of needs analyses as a basis for test development after having reviewed British discussions of the topic (eg. Weir 1983, Alderson

1988, Alderson and Clapham 1992), but nevertheless, he proposed a needs analysis research program to guide the development of the TOEFL 2000 test. The program was to be clearly guided by the purpose, ie. test development; it should implement a comprehensive and representative framework for data gathering; there should be triangulation in terms of sources and types of data and methods of data gathering; the concept of "need" should be defined comprehensively to include "wants" or "perceived needs"; and the aim should be to describe needs and necessary requirements at arrival to university, not the nature of skills which the learners will develop during their university course (Waters 1996:55-60). One of the recommendations that Waters made was that the needs of native English-speaking students should also be surveyed.

Ginther and Grant (1996) reviewed research on the academic needs of native English-speaking college students in the United States. They stated in their introduction that since the majority of the studies they reviewed served teaching needs rather than assessment ones, the direct usefulness of their results for test development would be limited (Ginther and Grant 1996:1). Rather, the results may clarify what TOEFL 2000 cannot do, ie. implement very extended tasks such as assignment writing, which was what many authors on composition and communication researched and discussed.

In a similar discursive-critical spirit, Ginther and Grant (1996:25-30) presented much more scepticism and criticism towards needs analyses as a basis for test design than Waters (1996). They concluded that tasks are very difficult to specify because there are so many levels to which individual judges can pay attention. For summary writing, for example, descriptors would have to include cognitive abilities, processing strategies, nature of source text, and variability in student interaction with text. They also pointed out that no more should be required of non-native college applicants than of native speakers, ie. the test tasks should represent college requirements prior to college-based instruction. They strongly recommended that the TOEFL 2000 program should begin not from analysing a broad array of needs but from the definition of the domain of the test. They suggested that since knowledge of English was the factor that most clearly differentiated between native and non-native students, this was what TOEFL 2000 should focus on. Further, they recommended that the TOEFL program should clearly state its rationale and allow this to guide data collection rather than let data guide the rationale.

The implications for construct definition from these studies were divergent. Waters (1996) suggested that task categorisation on the basis of

an analysis of contextual variables would be possible if research were conducted in a systematic way, and he proposed a research agenda for doing so. Ginther and Grant (1996) suggested that a better basis for the categories of test construction would be the nature of language knowledge. Thus, they preferred the starting point to be an individual's language ability rather than the context of use. Both proposals would lead to verbal descriptions of ability in the first instance but both entail categorisation, which may enable quantification at later stages of test development.

8.3.2 Construct definition: theoretical background

Four papers have been published in the TOEFL Monograph Series on the theoretical background work on construct definition that has been done in the TOEFL 2000 project. One of these, Chapelle et al. (1997), defines the theoretical framework that was developed for the project. Since this model was developed by the TOEFL Committee of Examiners (COE), it is called the COE Model. Three skill-specific papers discuss communicative language ability in a coherent spirit, and most of them refer to the COE Model. Hudson (1996) discusses reading, Hamp-Lyons and Kroll writing, and Douglas (1997) speaking. No comparable paper on listening has been published in the Monograph series.

Chapelle et al. (1997) presented the COE Model, which defines communicative language use in academic contexts. The model was based on a broad range of current theoretical concepts in applied linguistics all carefully referenced, although the writers particularly emphasized the model's relationship with Hymes (1971), Canale and Swain (1980), and Bachman (1990). The model specifies the components that the COE believes to be relevant in language use, as well as hypothesized the relations between the components. It was used in the project "to focus discussion on how to define what TOEFL 2000 is intended to measure" (Chapelle et al. 1997:1), with implications for both test development and validation.

The COE Model identifies aspects of the context of language use on the one hand, and hypothesized capacities of the language user on the other. It represented a summary of "existing research and current assumptions by researchers in cognitive psychology, applied linguistics, and language testing" because it was important that the model was up to date, and it identified aspects in both because the model must be useful for test development and validation (Chapelle et al. 1997:2). In the model, the context variable is divided into two interdependent entities, situation and performance. The situation in the case of TOEFL 2000 is academic, eg. a lecture or an office appointment, and the Model characterises those features

of it which were expected to influence academic language use, namely setting (ie. physical location), participants (ie. the individuals and their roles), task (ie. a piece of work or an activity with a specified goal), text (ie. type of language used to complete a task), and topic (ie. the content information that is being addressed) (Chapelle et al. 1997:6-9). The Model sees performance as the language user's contribution to the context. It consists of the linguistic or behavioural output of the language user in interaction with the situation. The Model does not claim that the aspects identified in the "context" are new, and indeed the sources of the categories are carefully referenced. The novelty in the TOEFL 2000 approach comes from two sources, firstly a commitment to the modelling of the individual's abilities in interaction with the context, and secondly from their operationalization of the properties of tasks and abilities in detailed content indicators, as will be discussed below.

The language user's abilities are seen as interrelated internal operations in the COE Model, ie. they are viewed as processing concepts. The processes begin with internal goal setting, which is motivated by the individual's perceptions of and responses to Context. The internal operations involve verbal working memory, which includes the joint interactions of a verbal processing component with metacognitive and online processing, language competence with linguistic, discourse and sociolinguistic knowledge, and world knowledge. The result of this interactive processing is internal processing output, which is the language user's representation of the situation or activity "so far" and which can lead to overt performance in terms of words and actions. (Chapelle et al. 1997:10-17.) These components are considered to interact in any language use situation, and the organising principle for their interaction is the situational context rather than the skills of listening, speaking, reading, or writing. This is why the Situation was selected as the basic unit of context that the test developers set out to define. This also explains why the TOEFL 2000 project would like to implement tests of integrated skills in situation-based tasks or test sections. However, the project also intends to test skills individually, and their "interaction" refers to all kinds of interactions, including examinees' interactions with texts and figures, ie. not only human interaction. Theoretically, the intentions of the project make good sense. It will be interesting to see, however, how these desirable developments can be implemented in actual tests.

As far as the definition of the construct assessed in TOEFL 2000 is concerned, the COE Model does not directly define it but rather it guides the test developers' decisions on what to define (Chapelle et al. 1997:21).

The model identifies situational and internal components in communicative competence, and because it incorporates the belief that the internal components are interrelated and activated by various features of the environment that surrounds a language user, it suggests that test development in TOEFL 2000 should begin by an examination of the types of academic contexts in which language is used. The test developers should identify key situations and hypothesize the abilities required by them in terms of goal setting, language processing, and linguistic, sociolinguistic and discourse competence. They should then construct relevant task formats and develop a scoring rubric for these (Chapelle et al. 1997:21-25). Whether this means that goal setting, language processing, and linguistic, sociolinguistic and discourse competence should then figure in the scoring and score reporting mechanisms is not specified by the COE Model.

The implications of the COE Model for validation are complex and, in the spirit of Messick's (1989a) definition, they include eg. content validity, construct validity, and the social consequences of test use. The developers' work on the construct definition, as evidenced in the COE Model and related TOEFL publications, reflects a commitment to forming a detailed understanding of the content of the test and the construct(s) that are invoked in the interpretation of its scores. In terms of consequences, Chapelle et al. (1997:35-37) discuss the need for evidence concerning the relevance and utility of the new test's scores, the value implications of score interpretation that reflect both the values of the theory underlying the test and the values it invokes in score users, and the social consequences of the introduction of a new TOEFL. The latter are elaborated further by Bailey (1997), and will be discussed later in this case report.

Hudson (1996) analysed academic reading in a second language in terms of "academic literacy", where the reader is both a cognitive information processor and an actor in a social environment. Hudson stated that, in general, "success or failure in reading performance can be addressed in terms of the interactions between the reader's (a) *automaticity*, the extent to which the performance of procedures no longer requires large amounts of attention; (b) *content and formal schemata*, the reader's mental representations of facts and skills; (c) *strategies and metacognitive skills*, the reader's strategies for monitoring the selection and application of actions; (d) *purpose*, the goal striven for by the reader, and (e) *context*, the interactional environment where the reading activity takes place" (Hudson 1996:4). The author reviewed existing reading research under these headings and listed implications for the TOEFL 2000 program in terms of Messick's

(1989a) validity framework, because he explained that this is the program's selected validation guideline (Hudson 1996:9).

The first implication that Hudson (1996:9-12) listed was that, to represent the construct adequately, the TOEFL 2000 program should expand beyond selected response tasks. However, rather than including constructed response tasks only, Hudson advised a balance between the two task types. Furthermore, he recommended the inclusion of authentic, situated tasks where skill modalities may be mixed, as well as the inclusion of thematically linked literacy sections in the test. The implementation of these recommendations, Hudson suggested, would support the test's representativeness of real world skills and avoid negative value implications that are associated with multiple-choice tests, while the inclusion of some selected response tasks would enhance the generalizability of the results. Through this recommendation, Hudson expressed his own values, but at the same time he recognised the value-ladenness of the new test. Hudson's last two implications were related to the conceptualisation and reporting of the test scores. To achieve relevance and utility, he recommended that, in addition to possibly providing a reading score, the certificates could also report a separate combined-skills score that could be labelled a literacy score. This would be based on the examinee's performance on tasks that integrated skill modalities. Reading could be combined with any other skill modality according to Hudson, and he provided an example which integrated reading, summarising, listening, and responding to questions.

The breadth of Hudson's concept of academic literacy was consistent with the COE Model, and his review of existing research was substantive. His discussion clearly focused on implications for testing, but from the perspective of theoretical literature rather than testing applications. Thus, Hudson did not consider examples of tests which have implemented, or have tried to implement, some of his recommendations. Examples such as the use of thematic linking in the Carleton Academic English Language test (eg. Fox, Pychyl and Zumbo 1993) or the attempt to develop integrated reading-writing tests in the IELTS (Hamp-Lyons and Clapham 1997) could have been used. The concentration on theoretical background led Hudson to raise, but not answer, the questions of score reporting, test equation, comparability of test forms, fairness, reliability, test security, and negative effects due to subjectivity. Nevertheless, Hudson's (1996) paper made progress in the concretisation of the definitions of the COE Model into actual test forms in the context of reading and literacy.

Hamp-Lyons and Kroll (1997) analysed academic writing from the perspective of current theories in composition in the light of the COE Model

of communicative competence. They proposed that writing in TOEFL 2000 should be considered discourse competence, that is, writing as an act that takes place within a context, accomplishes a particular purpose, and is appropriately shaped for its intended audience. The developers must consider the skills that are needed to succeed in an academic context and decide whether these skills are the same for all potential test takers, especially undergraduate and postgraduate students. To develop assessment principles, they must also consider whether the performance expectations are the same for both groups, find appropriate assessment criteria that do not disadvantage identifiable groups of test takers and score users, and balance the forward-looking perspective of academic needs with the experience of TOEFL 2000 test takers, who are not yet members of the North American academic community.

Hamp-Lyons and Kroll (1997:18-20) considered different theoretical approaches to writing assessment and argued that the most feasible approach would be to include more than one writing prompt and possibly allow test takers some choice in prompt selection. They pointed out, however, that little is known about writers as test takers, and that it is particularly important to investigate whether weaker writers are disadvantaged by the inclusion of choice in the test. They then considered the application of their observations into test specifications and tests for TOEFL 2000. Regarding prompt development, they raised the complex concepts of difficulty and accessibility and the adjacent implication that to create somewhat equal test forms, the test specifications must define prompt characteristics in a very detailed manner. When they considered test time, they concluded that studies showed that advanced writers were advantaged by increased writing time which would support planning and re-writing. In terms of scoring, they raised the possibility of using multiple-facet IRT analysis to model examinee ability, prompt difficulty, and rater harshness. Furthermore, they raised the possibility of reporting scores in a profile format where aspects of writing could be rated because this would give more detailed information on score meaning. They pointed out that the scoring criteria used by raters from different backgrounds should be studied further in the TOEFL context. Furthermore, Hamp-Lyons and Kroll (1997:31) pointed out that test development oriented research was needed on appropriate scoring criteria, rater training procedures, and validity.

Hamp-Lyons and Kroll (1997:32-33) closed their report with a consideration of costs, practicality, and washback from the test. They recognised the need to balance theoretical desirability with practicality, but their conclusion was nevertheless that direct assessment of writing through

multiple tasks was needed. They noted that the inclusion of productive writing tasks would increase costs, but they suggested that some costs might be won back through increased usefulness of the scores if they can give detailed information about the score holder's writing abilities and possibly provide actual samples of their writing. Thus, their report linked theoretical considerations with practical recommendations for test development.

Douglas (1997) provided a discussion of the theoretical background to testing speaking in academic contexts. After presenting a psycholinguistic speech production model and discussing its implications for assessing speaking, he discussed test methods as the test's representation of contextualisation cues in non-test language use and drew implications for possible formats of speaking tests and the rating of speaking performances.

Douglas's speech production model extends Levelt's (1989) and de Bot's (1992) language processing models. All the models identify internal processes of speech comprehension and speech production, which are considered to work simultaneously. The processes identified are conceptualization, formulation, articulation, auditory perception, and a speech comprehension. Douglas (1997) integrated this processing model with a communicative competence approach, and specified further especially the nature of the knowledge store through which the individual models the interactional context. Douglas distinguished two components in it, a Knowledge Component, which included world knowledge and language knowledge, and a Strategic Component, which included metacognitive strategies, language strategies, and cognitive strategies. Through these, the contributions of the Bachman (1990) view of language knowledge and the related COE Model views of strategic competence were related to the processing model. The processing model specifies that knowledge and strategic components feed material to the conceptualizer and thus influence message generation. This happens in two stages, macroplanning, which is conceptual, and microplanning, which is linguistic.

Douglas (1997) strongly emphasized the role of strategic ability for TOEFL 2000 because this is how the individual interprets the context. In the COE Model, the context is seen as one of the two central determiners of communicative competence. According to Douglas (1997:6-9), the strategic component includes two components. Metacognitive strategies include assessment of the context, goal setting, cognitive planning, and the control and execution of attention. Language strategies involve assessment of the discourse context, setting of communicative goals, linguistic planning, and control of linguistic execution. These are important because not enough is

known about them to interpret scores in terms of strategies, which are nevertheless expected to be centrally involved in performance.

For Douglas (1997), the implication of the processing model of speech production was that since test takers perceived context in a test situation in much the same way as they would in any other situation, the contextual cues, which in a test are embodied in the test method characteristics, must be carefully specified and rich enough so that all test takers interpret them in specified ways. This, in turn, would lead to comparability of scores. In accordance with the COE Model, Douglas (1997:12) connected contextualisation with the examinee's role interpretation of it and identified three levels: specific-purpose tests, generic-purpose tests, and "pan-technical" tests, which represent a middle level of contextual specification. In such tests, the context is defined situationally and linguistically but not to such a high degree that it would be a specific-purpose test. Douglas proposed that it was possible to define academic language use contexts in this way.

Douglas (1997:18-19) suggested that it was important to determine contextual variables carefully in a speaking test because variation in any contextual feature could lead to differences in examinee perceptions of the situation with consequent differences in performance. He also argued, coherently with the COE Model, that instead of attempting to minimise the effects of context on performance, testers should capitalize on their ability to control test method characteristics and use this in contextual manipulation and in score interpretation. He postulated a threshold of authenticity that would be necessary for examinees to perceive the situation as authentic, and suggested that context-based tests may be particularly relevant for tests of communicative language ability.

In terms of assessment, Douglas (1997:25-26) proposed that listening and speaking were so integrated as constructs that they should be tested together and that the reported score should also be a combined aural/oral score rather than two different scores. His other recommendations for TOEFL 2000 included the use of fuzzy logic in the scoring of performances and the use of computer rating to the extent possible. Douglas's (1997) paper thus presented a wide range of hypotheses for possible future developments, and a basis for construct definition in individual processing. He did not consider conceptual categories for score reporting in detail, so it is not possible to say whether he considered processing and contextual variables as relevant concepts for score interpretation.

The three papers discussed above all defined a skill-based construct, but in quite different ways. The complementary perspectives undoubtedly

helped the COE to consider different dimensions in the construct of communicative competence. However, as Jamieson et al. (2000:5-6) explained, when prototyping teams attempted to operationalize the constructs identified in the papers, they ended up having quite varied results. The modules "included extended reading and listening passages that were contextualized, were linked thematically, and contained integrated, performance-based writing and speaking tasks" (Jamieson et al. 2000:5). An evaluation of the modules revealed that the theoretical frameworks were not specific enough to allow systematic implementation in a test. Hence, more concrete frameworks for test specifications were needed. These will be discussed in the next section.

8.3.3 Construct definition: frameworks for test development

To address the needs of concrete test development, the TOEFL 2000 project wrote a framework document for the whole project and created skill-specific frameworks for reading, writing, listening, and speaking. The overall framework (Jamieson et al. 2000) reported the rationale followed in the development of TOEFL 2000 and defined a series of steps through which the developers proceed from the purpose of the test to construct-related variables that can be used to create test specifications and scoring criteria. The skill-specific frameworks followed the overall pattern to define their respective domains and characterise possible test tasks. They also considered operational constraints, outlined a research agenda to refine and validate each framework, and presented criteria for the evaluation of the new test against the existing one. Since all the skill-specific frameworks follow the same organization, I will discuss one in more detail and another to demonstrate how the specific frameworks differ. The detailed discussion will concern reading (Enright et al. 2000) while the comparison will be with speaking (Butler et al. 2000). These are also the two skill-specific frameworks officially published to date.

Jamieson et al. (2000:7-8) explained that the main motivation for the creation of the *TOEFL 2000 Framework* was improved measurement. However, there were also other potential benefits. The framework would provide a common language for the discussion of the test and the construct intended to be measured. Discussion about it would help build consensus about the measurement intent among those involved in the development. It would set parameters for task construction and score interpretation, while its detailed categories would support greater and better controlled construct representativeness. It would be possible to develop detailed specifications and possibly reduce test development costs. The detailed information about

the principles on which the test tasks and assessment criteria were built would enhance the construct validity of the test. Standards would be easier to set, and score reporting could be based on detailed empirical information about what was assessed in the test. The links between research, testing, practice, and public policy would promote the continued development of the test and an understanding of what the test is measuring. The list includes a lot of potentials which call for verification, but in principle the logic is solid. The basis of these benefits is joint concentration on the theoretical and measurement perspectives on score meaning.

Jamieson et al. (2000:10) began the presentation of the framework from a statement of purpose for TOEFL 2000. This is as follows:

> The purpose of the TOEFL 2000 test will be to measure the communicative language ability of people whose first language is not English. It will measure examinees' English-language proficiency in situations and tasks reflective of university life in North America. Abilities will be reported along one or more scales characterized by increasing levels of proficiency. Scale scores are designed to be used as one criterion in decision making for undergraduate and graduate admissions. Information derived from the proficiency levels may also be used in guiding English-language instruction, placement decisions, and awarding certification.

The statement implies that the framework must be anchored in theories of communicative competence, as discussed in the COE Model. Accordingly, an important basis of organisation in the test construct is situations in university life in North America. However, following score users' wishes, the decision was made to report scores on the four skills although they may be tested both independently and integratively (Jamieson et al. 2000:11-12). The main purpose of score use is the same as the current test's, although the purposes may be extended.

The major contribution of the *TOEFL 2000 Framework* is the connection it builds between the theoretical construct definitions and their operationalization in test tasks. This is done by the identification of the variables that define the task characteristics, the quantification of these variables, and the establishment of the connections between the variables, empirical item difficulties, and interpretive schemes that are built for TOEFL 2000 scores. Jamieson et al. (2000:24) noted that while work existed where authors such as Bachman and Palmer (1996) and Duran et al. (1985) had identified variables that influence communicative language ability, these had not been combined with a similar identification of the variables for language tasks or specified levels of ability, which they intended to do. They followed examples from adult literacy, where a set of validation studies on

the variables had been conducted by Kirsch, Jungeblut, Mosentahl and colleagues (eg. Kirsch and Jungeblut 1992).

The framework of the variables identified for TOEFL 2000 is complex, but because it is central to forming an understanding of the way in which the construct-task connections are made in the project, it is summarized below. A brief example of quantification of the variables will also be provided.

In the *TOEFL 2000 Framework*, the variables are defined for situations, text materials, and test rubrics. The situation variables that are defined are participants, content, setting, purpose, and register. These are considered to be "simultaneously at play in language tasks" (Jamieson et al. 2000:16). The text material variables relate to both task material and language produced by the examinee. The variables identified are grammatical features, which relate to the structure of sentences and vocabulary, pragmatic features, which relate to the intent of the text's creator, and discourse features, which relate to the nature and structure of the text as a whole. The discourse variables identified are rhetorical properties, which encompass definition, description, classification, illustration, cause/effect, problem/solution, comparison/contrast, regulation, or analysis, and text structure properties, which are defined separately for documents, prose, and interactions in the oral mode. Documents are defined as "written texts that consist of words, phrases, and/or diagrams and pictures organized typographically" (Jamieson et al. 2000:19). The test rubrics are defined by three sets of variables. These concern the questions or directives posed, response formats, and rules for scoring. The definition of questions and directives is most detailed, involving three categories: firstly, the type of information requested, ranging from concrete to abstract, secondly, the type of match, which refers to the way in which examinees must process text to provide the response, and thirdly, the plausibility of distractors, which concerns "the extent to which information in the text shares one or more features with the information requested in the question but does not fully satisfy what has been requested" (Jamieson et al. 2000:22). These variables are particularly important because they specify not just the textual characteristics of texts and items but the operations that the examinees must perform to answer the items successfully.

Jamieson et al. (2000:21) provide a simple example of quantification of the variable "type of information requested" in a test question. In a study of adults' and children's literacy, this was scored on a 5-point scale. A score of 1 was awarded to questions that requested concrete information, eg. the identification of a person, an animal, or a thing. A score of 3 was

awarded to questions that required the identification of goals, conditions, or purposes. A score of 5 was awarded to questions "that required examinees to identify an 'equivalent'", ie. there was an unfamiliar term or phrase and the examinees had to infer a definition, interpretation or predicting condition for it from the text.

In sum, this procedure enables test developers to express the content of an item through numbers. If several content characteristics are defined, combinations of "content numbers" can be created. These can be compared with observed item difficulties to see how well they account for observed variance in difficulty. When this was done with 20 reading and 20 listening items from the experimental computer-based TOEFL test, the three question/prompt variables of type of information, type of match, and plausibility of distractors accounted for 86 percent of the variance in task difficulty of the reading tasks, and for 79 percent of the variance of task difficulty of the listening tasks (Jamieson et al. 2000:28-29). The high explanation percentages were encouraging, although the difference between the skills indicated that some skill-specific modification of predicting variables or their quantification may be required. Furthermore, the categorisation of test items in this fashion might facilitate the explanation of why certain groups of items cluster at different regions of a scale of empirical difficulty – or, conversely, the explanation of constructs assessed in items of varying difficulty. When this was studied in the adult literacy surveys referred to above using a process that combined content analysis of the items, quantification of content variables, and the derivation of a set of rules to categorise the items (see Jamieson et al. 2000:31-37 for a summary description, and McNamara 1996 or the discussion of his examples in Chapter 4 for a related example), the results indicated that it was indeed possible to characterise the nature of items at five levels of difficulty. This enabled the creation of a construct-based scale that described prose task difficulty and examinee proficiency. The descriptor for Level 1 provides an illustrative example:

> Level 1. Most of the tasks in this level require readers to identify information which is quite concrete, including a person, place, or thing, or an attribute, amount, type, temporal, action, procedure, or location. To complete these tasks, readers must process relatively short text to locate a single piece of information which is identical to (or synonymous with) the information given in the question or directive. If distractors appear in the text, they tend to be located in a paragraph other than the one in which the correct answer occurs. Jamieson *et al.* (2000:36).

This is the type of scale that TOEFL 2000 desires to develop. However, the developers recognise that to be able to do this, the variables identified in the test and its tasks must be validated along the lines of the trial study with the

current TOEFL that was mentioned above. Furthermore, skill-specific frameworks for the new TOEFL test must be developed so that prototype tasks can be constructed and research conducted on them. Jamieson et al. (2000:39) conclude the *Framework* with a reference to the work of four working teams charged with "(a) using the current framework to operationalize specific frameworks for reading, writing, listening, and speaking, and (b) developing a research agenda to support the framework."

Accordingly, the *Reading Framework* (Enright, Grabe, Koda, Mosentahl, Mulcahy-Ernt and Schedl, 2000) applied the general *TOEFL 2000 Framework* and the developers' understanding of theoretical considerations in reading to develop a design for the TOEFL 2000 reading test. The authors identified three perspectives in reading literature, the processing perspective, the task perspective and the reader purpose perspective, and presented the current state of research in each of these areas in concise summary terms. They proposed that reader purpose offered the most practical tool to "explain the principles driving test design and test development to the general public" (Enright et al. 2000:5). This was in slight contrast to the task-based emphasis of the overall *TOEFL 2000 Framework*, but the researchers argued that their perspective was most useful for score interpretation. They also argued that it was possible to connect reader purpose with processing and task-based views, which would further enhance the explanation of score meaning.

Enright and her colleagues distinguished four main reading purposes that are relevant for TOEFL 2000. These are reading for information, reading for basic comprehension, reading to learn, and reading to integrate information. The first two were covered by the current TOEFL test, and a main contribution of the new test would be the inclusion of reading to learn and reading to integrate information, both of which are relevant activities in an academic context. This would enhance the construct representativeness of the new test. In terms of reader tasks, reading to find information and reading for basic comprehension required the identification and interpretation of information. With reading to learn, the reader tasks entailed were summary, definition, description, elaboration, and illustration. Finally, when reading to integrate information, readers must compare, contrast or classify, establish the nature of a problem or propose an integrated solution, explain or justify a case, persuade, or possibly narrate. (Enright et al. 2000:30-35.)

The text types that Enright et al. (2000:20-23) proposed for inclusion in the new test were exposition, argumentation / persuasion / evaluation, and historical/biographical narration. The last refers to text formats that define a

setting and a chain of episodes to reach a goal or solve a problem. The same type of organization is also followed in literary texts, which are relevant in some academic settings, but Enright and colleagues considered these to be outside the TOEFL 2000 framework because literary works can build on cultural references and background knowledge to such a high extent that it might unfairly disadvantage test takers. As for text formats, they proposed that, in addition to expository texts, TOEFL 2000 should also include document-type texts which encompass tables, schedules, and graphs that include non-prose presentation of text to be understood. Such textual elements are frequently included in study materials, which makes them relevant for TOEFL 2000.

In keeping with the TOEFL 2000 Framework, Enright and colleagues defined the situation variables of the reading tasks in terms of participants, settings, content, communicative purpose, and register, and characterised the text material in terms of possible grammatical/discourse features and pragmatic/rhetorical features. The organizing principle that they used to form a test rubric out of them incorporated the four reading purposes discussed above. Thus, for example, tasks that required reading to find information and reading for basic comprehension could be presented on texts characterized by various rhetorical patterns such as definition / description / elaboration, comparison / contrast, or problem / solution. According to their framework, the material should include both prose texts, documents (as defined above), and quantitative texts such as bar or pie charts. The task types that they proposed included multiple choice, point and click on text, point and click and drag, and open-ended response with words, phrases or sentences.

For reading-to-learn items, the authors proposed research to discover effective item types. Such research could include studies of timed reading where text material would be taken away from test takers after a set period of time and they would be required to summarise the text to show what they have "learned", and contrastive research without text removal. They also proposed exploration of short answer and selected response tasks as well as exploration with metatextual tasks. An item like this might require an examinee to say that a text employed a comparison, or sequencing by importance. Finally, they proposed the study of cross-modality integrating tasks; so that, for instance, a reading text and a listening text would provide complementary information for a writing task. These could also be used in the assessment of the last reading purpose that they identify, reading to integrate information.

As for task sequencing, Enright et al. (2000:38) proposed that the examinees should first be presented with a range of text material and a set of information location and basic comprehension tasks in multiple-choice and open-ended machine scorable formats. This would be followed by reading to learn and reading to integrate tasks which might include extended responses and a combination of modalities as information sources. They suggested that, in addition to reading purpose and item format, task difficulty could be manipulated through textual variables such as vocabulary, syntactic complexity, transition markers, amount of text and time, and competing linguistic distractors in the text environment. They briefly considered technological alternatives (Enright et al. 2000:40-42), among which there might be the combination of electronic and hard copy materials and the combination of texts, charts and graphs, sound, and video. The considerations emphasized speed of access, flexibility, and ease of use.

The logical stages of test design that Enright et al. (2000:44-48) identified for test development in general and for the TOEFL 2000 reading test specifically are construct identification, prototyping, pilot testing, and field testing, but they also pointed out that development is cyclical and iterative. This conforms to the theories discussed in Chapter 2 and the steps of initial test development identified in the summary model in Chapter 5. They proposed a detailed research and development agenda with several activities for each of the stages. The categories that they proposed to be observed in studies related to pilot testing were user acceptance, concurrent validity, construct representation, impact of construct relevant and construct-irrelevant test-taker characteristics such as length of study or length of residence in the US on item performance, the performance of native speakers, and the factors that affect item difficulty. Later studies on larger samples could focus on scale descriptors for score reporting based on task analysis, normative information on native speaker performance, and construct validation research. This would include studies of convergent and discriminant validity, construct representation based on known task characteristics, changes of score patterns over time, and differences between subpopulations. Finally, they recorded the need to evaluate the appropriacy of different psychometric models for scoring and scaling items (Jamieson et al. 2000:47-48). The consequences of score use could only truly be investigated once the test was operational.

In their concluding statement, they considered the advantages of the TOEFL 2000 reading efforts over the existing reading test. The main contribution was considered to be the articulation of the construct tested

and the links between this definition and the test design (Enright et al. 2000:49). Through empirical links to item difficulty, this would enhance the meaningfulness of the scores. The reading purpose framework would also expand the construct measured to reading to learn and reading to integrate information, which might make it possible for the test to discriminate better at higher proficiency levels. The improvements had the potential to build positive washback. If the research agenda proposed in the document is implemented, evidence will be provided about whether the improvements are in fact realized.

Butler, Eignor, Jones, McNamara, and Suomi (2000) presented an initial framework for research and development for the speaking component of the TOEFL 2000 test. Their discussion followed the same pattern as Enright et al. (2000) in that they reviewed existing literature, presented a speaking framework where the task characteristics and variables for quantifying them were identified, discussed technological issues in test implementation, and presented a research agenda. Their treatment differed slightly from that of Enright et al. (2000) on two accounts. Firstly, there was not as much scaffolding in existing research on speaking to support the creation of a detailed assessment framework as there had been for reading, and therefore the research agenda proposed began with a call for literature reviews on speaking needs, register, oral genres, cognitive demands of tasks, and existing speaking tests. Secondly, since speaking was a productive skill, the assessment considerations that test development entailed were complex, and furthermore they presented technological challenges that were unlikely to be met by machine scoring in the very near future. Thirdly, large-scale assessment was a particular problem for an interactive skill because operational constraints meant that the TOEFL 2000 speaking test would likely be indirect, most likely computer-mediated, rather than direct interaction with one or more speakers. The operationalization of even an indirect test posed technological challenges that would require further investigation.

In spite of the early stage of development, Butler et al. (2000:4-8) attempted to characterise the test domain through the categories of the *TOEFL 2000 Framework*. They used the situational features of participant roles and relationships, topics, settings, and register, but although they presumed that some features would be related to task difficulty, they did not express the hypotheses in specific quantifications. The stage of test development was probably too early to justify this. They discussed the discourse features of the speaking test in terms of genre and pragmatic features, on the one hand, and structural features, on the other, but again did

not quantify the dimensions and called for research in these areas. Their consideration of test rubric included a discussion of types of response quality of performance. The response types of reading aloud, elicited sentence repetition, and constructed response were considered, and although preference was expressed for constructed response, the first two alternatives were not excluded because in spite of their construct limitations they allowed machine scoring. As for quality of performance, the authors preferred polytomous scoring where different levels of performance could be distinguished. They expressed a preference for analytic scoring over the holistic scoring employed in the current Test of Spoken English, because it would potentially provide more detailed score reporting and clear differentiation between tasks in terms of their difficulty.

For dimensions of task difficulty, Butler et al. (2000:16-18) considered factors defined in the work by Skehan and colleagues (eg. Skehan 1998). Skehan's task conditions, which were discussed in Chapter 4 above, included continua such as small to large number of participants and/or number of elements discussed, concrete to abstract information and task, here-and-now versus there-and-then information, and familiar versus unfamiliar information. Butler et al. also considered other factors proposed by Skehan (1998) that influenced task difficulty, such as time pressure, opportunity to control task, surprise elements, and visual support. Furthermore, they noted that Skehan (1998) had raised the notion of learners' attentional resources and their influence on the accuracy, complexity, and fluency of learner performance. In their research agenda, Butler et al. (2000:21-22) called for research on the validity of the variables identified. The same concerned the variables identified through the *TOEFL 2000 Framework*, especially structural features of discourse. Furthermore, Butler et al. called for initial investigations of operationalization, including research in Natural Language Processing, which might enable the creation of automatic scoring mechanisms.

8.3.4 Construct definition: measurement implications

Since the TOEFL 2000 project is not at the prototype trials stage yet, it has not been possible for the team's psychometricians to conduct empirical analyses of the psychometric properties of actual tests. However, some preparatory work has been done, and I will briefly discuss two papers with direct relevance to this aspect of the project. Carey (1996) discussed the psychometric issues raised in contexts where performance assessments had been used in high stakes tests. The motivation for this paper was the plans that TOEFL 2000 would include performance assessments and tasks that

integrated skill modalities. Tang and Eignor (1996) investigated empirically the feasibility of simultaneously calibrating dichotomously scored (right-wrong scored) and polytomously scored (scored using a multi-point scale, eg. 1-5) TOEFL items. The motivation for this paper was the likelihood that TOEFL 2000 would include items of both types.

Carey (1996) organised her discussion of psychometric concerns with performance assessments according to Messick's (1989a) validity categories of test-based evidence and testing consequences. Following Messick's (1994) statement that the validity criteria for traditional and alternative assessments are essentially the same, she listed the validity concerns under four questions: whether the intended domain is assessed through the performance tasks, whether the domain is well sampled through the range of tasks, whether it is possible to draw inferences from test scores to the domain sampled, and what can be inferred in a diagnostic sense from a performance that is not appropriately high. She concluded that the most serious threats of performance assessments regarding these challenges were the task specificity of scores, found in several studies and the low number of tasks that could be administered because performance tasks are typically complex. This endangered the generalizability of the results. She also discussed the reliability threats of scoring, although she admitted that in some studies, careful development of scoring rubrics and training of raters had been found to improve the reliability of ratings. On the point of generalizability, she also noted results that an increase in the number of tasks administered led to greater increases in generalizability than an increase in the number of scorers did. Furthermore, Carey (1996:8-10) raised test equating as a serious problem if large proportions of the future examination were performance-based. Equating through traditional psychometric means is difficult if the number of items per test is small.

Of consequential concerns, Carey (1996:11-12) raised fairness, test preparation, test security, and legal issues. Possible bias would be an issue whether a test was performance-based or traditional, but the lower number of tasks that tests might contain with performance assessments made bias balancing more difficult to address. TOEFL 2000 might generally be expected to lead to positive washback, but Carey warned that the objectives of the test would have to be formulated clearly to support this. Test security was an issue because long, complex narrative tasks were memorable; if test tasks were leaked, some examinees could gain inappropriately high scores because they were familiar with a particular task, not because they had the skills required by the task. Finally, legal issues entailed by the test were

mostly due to the potential for bias against identifiable sub-populations and limited validity if small numbers of items were included.

Carey's (1996) treatment of the issues raised in high stakes testing contexts about the psychometric properties of performance assessments was quantitative in approach. The problems of generalizability and score variability were real and serious, especially if the construct was defined only or mainly in traditional quantitative terms. Carey (1996) did not consider the alternative challenge posed by the theorists who worked on proposals of test development, namely that new psychometric models might be needed to support the quality of the new test. Her discussion indicated that there were serious discrepancies between what the test developers wanted to do and what the psychometricians' current means enabled them to do.

Tang and Eignor (1997) addressed one of the issues raised by Carey (1996), that of test equating if TOEFL 2000 included both dichotomously and polytomously scored items and items from different skill modalities. With data from the existing TOEFL, TWE and TSE tests, they examined the feasibility of calibrating combined skills tests with reading and writing, on the one hand, and listening and speaking, on the other. This would imply the reporting of a combined reading-writing score, on the one hand, and a combined listening-speaking score, on the other. The study was also method-technical in that the researchers sought for a practicable way of calibrating items that used the two scoring types simultaneously.

With data from 1500 test takers for three reading-writing combinations and from 434 and 502 examinees for two listening-speaking combinations, the results indicated that it was psychometrically possible to make the two skill modality combinations where both dichotomously and polytomously scored items were included, since the score distributions for all five combinations yielded sufficiently unidimensional results. A principal component analysis for all forms showed a dominant first factor that explained more than 40% of the score variance in all five cases (Tang and Eignor 1997:16). This was presumably considered an acceptable result, especially since integrated tasks were not included in the design, because the existing tests had been developed as skill-specific units. The researchers noted that the number of examinees for the listening-speaking combination was extremely low and should be increased in future experiments. The study of calibration methodologies indicated that a combination of the three parameter logistic model and either the generalized partial credit model or the graded response model could be used to analyse the data. The psychometric concerns were reported in detail including the equations used in various statistical models that the researchers applied on the data.

To a non-psychometrician, Tang and Eignor's (1997) analysis indicates that the psychometric implications of the desirable conceptual developments of TOEFL 2000 are more serious challenges than I would have expected. It may indeed be, as Jamieson et al. (2000:6) state, that the new test requires the development of new psychometric models to investigate their measurement quality. Whether the construct definition is simple or complex, it is crucial that it can be shown that the test gives consistent and predictable information about performances. The only widely recognised way to show such consistency used to be score distributions. The TOEFL approach where tasks are described through quantified variables that relate to underlying theoretical concepts, on the one hand, and that can be related to measurement indicators, on the other, provides a starting point for a different dialogue. The approach has already tightened the description of the abilities assessed in the test. If connections between item properties and score variations can be made and if a suitable psychometric approach to prove the measurement quality of the new test can be found, it should enable the delivery of more detailed score information. This should also make it possible to say in more detail what was measured in the test and what was *not* measured. Continued development work on TOEFL 2000 may show whether the calibration techniques trialled by Tang and Eignor will be combined with other measurement information for the test or whether indeed new measurement models need to be found.

8.3.5 Validation work

The studies that I have discussed in the TOEFL 2000 case above deal with construct definition and in a sense they could all be categorised as part of validation. Specific sections of the papers that I have discussed directly mention validity implications. The presentation is most detailed in Chapelle et al.'s (1997:29-37) treatment of the COE Model's implications for validation and Enright et al.'s (2000:43-50) research agenda for the reading section, which I will discuss below. Moreover, there is one TOEFL Monograph (Bailey 1997) that considers the social implications of the introduction of a new TOEFL test on those who are affected by it and on language teaching curricula more widely. It provides evidence of the project's commitment to the consideration of the consequences of test use.

Chapelle et al. (1997:29-30) defined validation in accordance with Messick (1989a). This view was discussed in Chapter 3 above, and in broad terms, it entails evidence and arguments that justify the outcomes of testing. The real world outcomes of TOEFL 2000 are that scores from the

test will be interpreted as indicators of an examinee's communicative competence in English in academic contexts and that the test and its scores will be used as one criterion when admissions decisions in universities are made. The latter, as a high stakes decision, means that learners will want to do well on the test, so it will be in the interest of providers of teaching to offer language education that makes this possible. In this way, TOEFL 2000 can be expected to have some influence on the way that English is taught at least to certain groups of students. According to Chapelle et al. (1997:29-30), the test developers could produce justification for these outcomes from the evidence of the construct validity of the scores and of the relevance and utility of the test, and from studies of the value implications of score interpretations and of the social consequences of the introduction of the test. To do this, baseline data about the impact of the current TOEFL test would be needed. Chapelle et al. (2000:30-36) discussed the kinds of evidence that the project would need to develop to support the validity of test introduction and score use. Concerning construct interpretation, these included evidence of the relevance and representativeness of test content, results of empirical item and task analyses to understand what is being tested, studies of the internal structure of the test in terms of relationships between items and test sections, studies of the external structure of the test through correlational evidence with other measures, and experimental manipulations to see if the scores vary in predicted ways if the test is changed in a controlled fashion or if groups of examinees take the same test under different conditions. Evidence of the consequences would entail studies of the relevance and utility of the scores and studies of possible unintended consequences, studies of value implications such as the privileged status of particular varieties of English, and studies of the social consequences of the introduction of the test. Chapelle et al. (1997:37) concluded with the statement that the validation work related to TOEFL 2000 would not be simple, but that it also held potential for the future development of validity inquiry as the results might show that the ideals cannot be realized and alternative practical validity criteria might have to be found.

Enright et al. (2000:43) similarly began their validity consideration from Messick's (1989a) definition and pointed out that of its six perspectives, substance and consequences had received increased emphasis in recent years. The substantive aspect was concerned with construct representation, in other words, the kind of work that had been conducted in the TOEFL 2000 project to define what would be assessed and develop empirical links to show that this was done. The link in their case especially

concerned explanation of task difficulty and its psychological counterpart of examinee ability. Enright et al. (2000:43) noted that this followed Messick's (1995) distinction of two dimensions in construct representation, domain representativeness in terms of content and processes and psychological representativeness in terms of features, knowledge, and processes involved in completing a task. The consequential aspect focused on score use and its intended and unintended consequences. For TOEFL 2000, the potential for positive washback was one of the motivations for the revision of the test. One of the possible unintended negative consequences might be reduced access to higher education because of increased costs of the test (Enright et al. 2000:43).

In accordance with modern validity views, Enright et al. (2000:43-44) contended that validation should be an integral component in all the stages of test design, which they listed as construct identification, prototyping, pilot testing, and field testing. The types of evidence and issues that would have to be addressed were operational, psychometric, and construct-related, and different kinds of validity evidence would become available at different stages of development. This was constrained in particular by the number of task exemplars that existed at any stage and the number of participants that the developers wanted to involve in the evaluation of the tasks. Thus, construct-related considerations would be possible at all stages, operational considerations would need to be considered early but could only be evaluated once a range of task exemplars existed, and psychometric issues could be addressed empirically only after large groups of examinees had taken part in prototype tests (Enright et al. 2000:44-48). Consequences could only be considered empirically after the test had been in use for some time.

Bailey (1999) summarised the existing literature on second language testing washback and drew implications for how the impact of TOEFL 2000 might be investigated. Bailey used Hughes's (1993; in Bailey 1999) distinctions between participants, processes and products in the teaching event to clarify her discussion of where the effects of washback might be seen. She identified students, teachers, materials writers, curriculum designers, and researchers as possible participants whose perceptions might be affected by washback. In addition to perceptions, washback might influence the processes in which these people are engaged, and the products of their learning, teaching, materials writing, and/or research processes. All of these, in turn, might have an influence on the test.

Bailey (1999:26) concluded that teachers' perceptions are most often investigated when the effects of washback are analysed, and she suggested

that student perceptions should also be investigated in the future. Her study of the processes and products of teaching suggested that the role of textbooks, authors and publishers in inducing washback was considerable, and that more research was needed on this (Bailey 1999:35). As for implications for TOEFL 2000, Bailey proposed that a range of methods should be used and that triangulated data on the potential washback of the new test should be gathered. She proposed the use of well-designed observation procedures, interviews, questionnaires and discussions in a triangulated fashion to develop multidimensional data on the effects of TOEFL 2000 on language education. Further, she suggested that it would be important to control for *when* the data was gathered, near an examination date or a long time from it. She also suggested that learner self-assessments might be gathered together with TOEFL 2000 performances to investigate the relationship between the test and learner autonomy, which may be related to computer use. Lastly, she focused on the possible specific washback effects that might ensue from the fact that TOEFL 2000 was computer-mediated and employed adaptivity. Such effects might include that learners become more familiar with computers because they prepare for the computer-mediated TOEFL test.

8.4 Case summary

The TOEFL 2000 case concerned initial test development. The stages that the developers identified in it were construct identification, prototyping, pilot testing, and field testing, which were implemented in an iterative fashion (Enright et al. 2000:44). This coheres very well with the top half of the model of test development in Chapter 5. The development reports indicated that special emphasis was placed in the initial development on theoretical construct definition. This meant that the validation side of Figure 3 in Chapter 5 was also included in the development operations. The development reports showed that test development and validation were connected both at the theoretical level and in the development activities. This was because both test development and validation work were expressly guided by a construct rationale.

The developers of TOEFL 2000 defined validation according to Messick's (1989a) broad concept, which included considerations of score interpretation and score use in terms of construct evidence and analysis of value implications and the social consequences of test introduction and score use. Because the reported stages of the development did not yet encompass prototype evaluations, the actual validation efforts

concentrated on construct definition, especially the building of connections between theoretical concepts that defined measurement intent and the measurement properties of the test. The linking work involved the identification and quantification of theoretically based variables that were considered important objects of assessment. The same variables would be used in task specification and the development of scoring mechanisms for the test. Such detailed construct basis holds promise for providing empirical evidence for the content meaning of the scores. The development revealed that it was possible that new measurement models would be needed to accommodate this conceptually multidimensional basis for score interpretation. In relation to the other areas of validation identified in Figure 3 in Chapter 5, identification of plausible rival hypotheses did not yet figure in the publications. Future validation research was planned, including some plans for examination of consequences. However, no actual studies of impact were conducted, although Bailey's (1999) report made concrete suggestions about baseline studies that would be needed.

The construct definition in TOEFL 2000 is detailed and theoretically based. The variables that are used for defining the construct relate to examinee abilities, on the one hand, and the properties of the context of language use, on the other; in other words, the definition is interactionalist. Its development entails the description of task characteristics from three perspectives: the textual features of the task, the situation invoked in it, and the task rubric, which identifies the operations that are expected of the examinees in order to perform the task successfully. The textual variables include grammatical/discourse features and pragmatic/rhetorical features, the situation variables define participants, setting, content, communicative purpose, and register, and the rubric/operations variables identify examinee abilities that have been derived from existing theory and considered useful for explaining task difficulty in TOEFL 2000. The construct definitions, developed by modality-based teams, will be used in the development of prototype tasks and scoring rubrics for the test.

The development of TOEFL 2000 up to the current stage has shown that the practical work required when content and measurement concerns are combined in test development is very complex and demands a varied range of expertise. The development report on the Reading section showed that theory offered several possible perspectives, including processing, task, and reader purpose perspectives, among which the developers had to make a choice. They chose the reader purpose perspective because they considered it the most useful for score interpretation. Other practically motivated choice issues were involved in the development as well, for

example when the four skills were selected as categories of score reporting in response to score user wishes. Examination of practical implementability showed that the new developments were a challenge not only theoretically and psychometrically but also in terms of test delivery. Furthermore, most progress had been made in reading and even this test was not yet at the prototyping stage. The implications of the project's approach to test development for the other skill modalities remain open.

Judging by the studies conducted on the development and validation of TOEFL 2000 so far, the developers take the challenge of testing communicative competence very seriously. They are committed to construct representativeness, but at the same time they also call for psychometric indicators of measurement quality, and for the development of new psychometric indicators if the old means restrict the construct too far. This is a tall order, and it is difficult to say what the response will be. The development is in early stages as yet, and it may be a number of years before the ideas are implemented in a functional test. The practical solutions for the combination of theoretical and psychometric concerns in the measurement of complex constructs have not been worked out. The consequences of the introduction of the test still await their realization.

9 TEST DEVELOPMENT AND VALIDATION PRACTICE: CROSS-CASE ANALYSIS

In this chapter, I will answer the main research questions in the case study by considering the joint results of the case analyses reported in the previous three chapters. The questions were:

- What are the similarities and differences between initial and operational test development and validation?
- How does the nature of the construct definition in a test influence the rationale followed and the quality criteria used in test development?
- How does the nature of the construct definition in a test influence the rationale followed, the questions asked and the results published on validation?
- How do the examples of realised practice in language test development and validation correspond to recommendations from theory?

I will take up each question under its own sub-heading. In accordance with good case study practice (Yin 1994:149), I will then discuss alternative and additional perspectives which may explain some of the differences between the cases. Finally, I will close the chapter with a summary of the results.

9.1 Initial and operational test development

The analysis of test development in Chapter 2 and the summary model of stages of test development in Chapter 5 showed an expected difference in initial and operational test development. The main cause was the standardisation that was expected to happen in formal examinations when they are published. Before it, all the development and validation activities had the potential of changing the goals and procedures of the other activities, because the object was to develop as good a test as possible. After it, the construction of new test forms and the arrangements for test delivery begin to follow standardized procedures that are designed to enhance score comparability. This requires monitoring and maintenance. Furthermore, new activities must be set up to deal with score data from actual testing rounds.

The cases in the three previous chapters enabled the examination of the whole test development cycle in the light of two cases. The published reports reflected the predicted differences in two main areas. One was a

focus on the clarification and improvement of construct definition during initial development, and the other was the availability of score data for analyses of the measurement characteristics of the test. Discussions of the test construct were concentrated on initial test development, as is shown in the reports on IELTS and TOEFL 2000. No similar construct discussions of TOEFL Reading or the operational IELTS had been published. This is possibly because publicly available tests are commercial products. While the optimal nature of the construct assessed may continue to be investigated by the examination board, at least in the form of evaluations of its acceptability to stakeholders, it would not be good publicity for the examination to criticize its current construct if they could not show that efforts were being made to revise the test. As shown in both IELTS and TOEFL 2000, even if a revision is in progress, criticism of the current test must usually be inferred from the fact that a revision is in progress rather than from explicit critical reviews by the test developers. Given that the material was published reports, this does not mean that such reviews were not conducted, just that they were not published.

As concerns score data, the case reports showed that psychometric studies were not available on TOEFL 2000, which is only at its initial stage of development. In contrast, several psychometrically oriented studies were published on the operational TOEFL Reading test. The same was not true of IELTS, possibly for reasons associated with assessment cultures, to be discussed below.

9.2 The influence of the nature of construct definition on test development

The results from the three cases of test development and validation indicate that the relationship between construct definition and test development is complex. To compare the cases in detail, it is necessary to specify the dimensions of construct definition that are relevant and also consider where the official "construct definition" for a test can be found.

In the design of the case study, I distinguished between theoretical and psychometric construct definitions. I considered the TOEFL Reading case to represent a primarily psychometric construct definition, the IELTS case to represent a primarily theoretical construct definition, and the TOEFL 2000 case to represent the combination of both. On the level of publicity information for the tests this is the case. However, the specifications for neither TOEFL Reading nor the operational IELTS have been published. Yet the specifications are presumably the source that the test developers use

when they implement the construct in their test development work. The TOEFL 2000 case was not relevant to this concern because the project has not yet reached a stage of specifications and prototyping.

Because of this difficulty in determining exactly where the construct definition is presented, it could be stated that it is difficult to know whether the theoretical construct definition truly was less detailed for TOEFL Reading than for the other two. It is possible that a detailed content and construct definition of the TOEFL Reading test exists and that it is used as a guideline when the test is constructed; indeed the Chyn et al. (1995) study on the automated item selection procedure for test construction indicated that this may be the case. Nevertheless, the comment can be made that the developers of the TOEFL Reading test seem to feel some unease about the detailed theoretical construct definition since it has not been published or discussed in studies. This may be related to the results of psychometric analyses of TOEFL scores which indicate that the test is psychometrically unidimensional. Based on the evidence from IELTS (Hamp-Lyons and Clapham 1997), it is possible to say that content-based considerations of the construct measured were used to judge the comparability of experimental versions of the writing prompts. When the prompts were found widely different in terms of their functionality and linguistic demands, the solution was to make the construct definition more detailed so that future writing prompts would be more comparable. In the case of the IELTS grammar section (Alderson 1993), in contrast, the rationale used for its deletion from the test battery built on an argument of construct coverage, quantitative information for test reliability, relationships between section scores, and the practical argument that a shorter test would be more economical and thus more acceptable to the funders.

As for the influence of psychometric construct definitions on test development, it is possible to say that quantitative item and test information is used in the construction of operational TOEFL Reading tests and judgement of the quality of draft items. According to Peirce's (1992, 1994) and Chyn et al.'s (1995) reports, only items with acceptable psychometric properties are included in the operational item banks and only tests with acceptable test information functions are released for operational use. Comparable information for the IELTS test construction procedures has not been published, which makes comparison difficult. Mirroring the discussion on the verbal construct definition, it is possible that the IELTS developers feel some unease about the psychometric properties of the test because information on them has not been published. However, the study of pretest data (Griffin and Gillis 1997) indicates that some psychometric

analyses are conducted in IELTS, and data from such studies may be used in the construction of operational test forms.

In relation to the expectation that the reports on test development would show clear differences in terms of test development practices (see Chapter 5), the comparison across cases indicated that the expectation was not borne out. The reasons for this can be viewed in different ways. On the one hand, it was clear that the definitions of the variables on which I based the expectation were loose, because I had not considered the possibility that different published and unpublished construct definitions might exist and that these might influence the results. On the other hand, in a real-world setting such as the present case study, it is not possible to control variables in a similar way to experimental studies. From this perspective, it can be concluded that the unclear evidence lends weak support to the expectation. Published research reports showed that in the TOEFL Reading case, psychometric considerations were important in test development and the developers were confident enough about their rationale for using them to publish details about their use. Similarly, published research reports showed that in the case of initial development of IELTS, the theoretical construct definition was considered important in test development and the developers were confident enough about its importance and relevance for test development to publish a report about its use as a basis for at least one test development decision and its contribution to other development decisions.

9.3 The influence of the nature of construct definition on validation

The results of the case comparison in terms of validation work indicated that the nature of construct definition in each case had an influence on the questions asked, the materials studied, and the results published. None of the cases exclusively concentrated on either psychometric or theoretical validation questions, but differences of emphasis were found.

In the TOEFL Reading case, a clear majority of the validation studies were psychometrically oriented. They concerned the measurement properties of the scores and score dimensionality, relationships among section scores, and relationships between TOEFL and other tests. The results showed the favourable measurement qualities of the test. Validation studies that were motivated by construct description and the assessment of construct representativeness were also conducted (Duran et al. 1985, Bachman et al. 1995), but the studies made it clear that the analysis was based on a single test form, not an evaluation of the test specifications or test construction principles.

In the IELTS case, the majority of the initial validation studies emphasized theoretically motivated construct clarifications, although this is not a particularly strong argument because quantitative studies at the early stages of test development would be unlikely simply because the nature of the tasks can change in response to development findings, so that it does not make sense to administer early versions to large groups of participants to analyse scores. Nevertheless, the focus of the studies was the verification that the right and appropriate construct was measured in the test, and the researchers argued that the results benefited both test development and validation (eg. Alderson 1988). The validation studies on the operational IELTS investigated both theoretical and psychometric construct dimensions but score data was used in most studies whereas theoretical rationales provided the starting point in only a few studies. However, the sample sizes in most of the validation studies were small and reliability and errors of measurement were most often not addressed; Griffin and Gillis's (1997) analysis of the IELTS pretest data was an exception. Whereas this showed neglect of psychometric quality criteria expressed eg. in the *Standards* (AERA 1999), the theoretically motivated study of the authenticity of the Task 2 Writing items (Moore and Morton 1999) showed that construct issues were tackled with seriousness and the results published even if they indicated some room for improvement in the test.

In the case of TOEFL 2000, the development and validation work included both theoretical and psychometric concerns from very early in the process. Work was focused on the development of a detailed theoretical and measurement understanding of the nature of the construct to be assessed. Attempts to develop both verbal/theoretical and quantified definitions of constructs through task characteristics were made and ways were sought to verbalise the content meaning of ranges of item difficulty. The activities were not yet at a stage where the psychometric properties of prototype items could be investigated, but preparations for analytic procedures were made.

In the case study framework in Chapter 5, I discussed my expectation that validation studies would show the influence of verbal and numerical construct definitions less clearly than test development. The expectation was based on the contention that instructions for research rationales and procedures were clear for psychometrically motivated validation studies and not equally clear for theoretically motivated studies. In fact, validation studies on the three tests showed the influence of the type of construct definition employed more clearly than the reports of test development. Construct-based validation studies were conducted in the two cases where

the theoretical construct definition was emphasized regardless of lack of detailed examples from theory. The means chosen by both the IELTS and TOEFL 2000 developers was the characterisation of tasks, especially through task demands. The original motivation for the construct studies was probably theoretical. At least in the case of TOEFL 2000, new methods for construct description were sought because the project was committed to a construct validation rationale that stemmed from current validity theory. This emphasizes the grounding of validation work in a detailed definition of the construct assessed. What the test developers provided were examples of how such detailed definitions might be developed and verified.

9.4 Correspondence between theory and realised practice

The cases analysed in the three previous chapters cannot be considered to represent the whole range of existing test development and validation practice, but they brought up a range of examples. Because the cases were selected on the basis of differences in construct definition, their correspondence with recommendations from theory highlights differences in this regard. In the case study framework, the theoretical recommendations for test development and validation were summarised on the basis of discussions in Part One of the thesis in Figure 3 in Chapter 5.

In terms of test development, the correspondence between theory and the case reports of practice was seamless. The cases illustrated the complexity of language test development work and, especially where initial test development was concerned, provided evidence for the integrated and iterative nature of the activities. As regards operational test development, similar complexity was evident although it was also clear that one area of activity identified in the Figure 3, operational administration, was not a research concern whereas the development of new test forms, monitoring and maintenance of test quality, and empirical validation were. The reports on these activities followed the guidelines and examples provided in theoretical writing. The emphasis on psychometric properties was clearly stronger in the TOEFL Reading case than in IELTS, whereas construct-based studies of what was assessed in the test formed a more comprehensive part of the examination system in IELTS. This was reflected in the way in which the construct studies on IELTS focused on, and drew implications for, the whole system rather than individual test forms. The lack of publications on the reliability of operational IELTS is nevertheless a deficiency when compared with professional standards for test monitoring.

Where validation is concerned, the correspondence between the reported cases and recommendations from theory was not complete. The published work concentrated on theoretical and/or psychometric considerations of the construct, while some studies focused on proposed and specific contextualised uses of tests. Serious empirical studies of test impact were not conducted, however. Moreover, critical analyses of the tests and especially the values that guided test development were very rare. Part of the reason may have been the examination boards' wary attitude towards publishing critical studies, possibly because it might affect sales or even result in legal cases. Part of it may be that such value discussion, while recommended in theory for many other areas of social life as well, not just language testing, is actually quite difficult to implement in practice, both because the fora for such discussion are missing and because critical discussion is socially difficult. Empirical studies of impact may be missing also because the concept of impact is complex (see eg. Alderson and Wall 1993, Messick 1996, Bailey 1999) and its empirical implementation is thus both demanding and costly.

In terms of the area of validation that the test developers did address, construct concerns, the cases differed as to how central the theoretical construct definition was stated to be for the development project. Only the TOEFL 2000 case could be considered to follow the recommendations because of its commitment to both the psychometric and the theoretical defensibility of the assessment system. This is required by theory because the term *construct* invokes both dimensions. Chapelle (1998:33), for instance, defines constructs as meaningful interpretations of observed performance consistencies. Meaningfulness calls for theoretical backing and the proof of consistency calls for quality of measurement. In the IELTS case, work was being done to support the meaningfulness of the scores, while the publications about the TOEFL Reading case bore evidence to the developers' commitment to measurement quality. Nevertheless, both cases were one-sided when seen from the perspective of theory's recommendations for validation practice. The TOEFL 2000 evaluation must be followed by the caveat that the development is in early stages, but the reports on it published so far bode well for the continuation, provided the principles are upheld.

9.5 The influence of alternative and additional perspectives

In the case study design, I divided the three cases into different categories according to the nature of the construct definition in them, and in this

chapter I have analysed cross-case results in these terms. However, because the cases focused on real life practice rather than controlled experiments, there were other differences between them besides the nature of the construct definition. In this section, I will discuss some of the additional differences and assess their possible influence on my results.

9.5.1 Published and unpublished test development work

When I presented the case study design in Chapter 5, I listed a number of caveats. The most important among them was that the case material did not include observations of practice but only published reports. One implication of this was that I would be unable to analyse some work related to test development and validation because it was not a topic of publication. Another implication was the case made by Spolsky (1995) that test development decisions were often driven by institutional forces rather than theoretical or measurement-technical rationales. Thus I might be able to find *that* a decision was made, but not find the reasons *why* this was done. I argued in Chapter 5 that while the reliance on published documents limited what I was able to analyse, the published studies would show what the test developers or testing boards considered important concerns and defensible professional practice. Nevertheless, some actual practice in test development is likely to have been excluded.

The case material highlighted one particular category of material that was not published, namely test specifications. The operational specifications for TOEFL Reading and for IELTS were confidential to the testing boards. I was interested in them because, guided by test development theory, I believed that they contained a detailed construct definition for the test. In the analysis of the TOEFL Reading case, I found some references to skill dimensions or task properties that were attended in test construction and that were specified in the test specifications. However, no detailed presentation or discussion of the categories had been published. The publicity material on the TOEFL Reading test gave a brief definition of the construct assessed. In the IELTS case, discussions of the nature of the skills to be assessed were published at the initial stage of test development, even if operational specifications were confidential. The publicity material on the test made use of the descriptors discussed at the early stages of test development and possibly the existing specifications. My conclusion was that this showed a different emphasis on the type of construct definition that the test developers thought important. This is the only conclusion that I can draw; it is possible that detailed considerations of skills assessed were

important in the construction of TOEFL Reading tests, but they were not published.

One of the cases, IELTS, also revealed another kind of material that was not published. This was psychometric information on operational test forms. Especially when compared with TOEFL Reading, the amount of measurement information published on the IELTS was minimal. Similarly to the construct concerns, I must conclude that it was possible that the psychometric properties of operational IELTS test forms were monitored at the testing board. The fact that they were not published as a rule led to the conclusion that the measurement side of the construct definition was not considered important by the test developers.

The case material also probably illustrated Spolsky's (1995) point that institutional or political reasons motivated test changes. For instance, I could not account for how or why the decision was made to change the score reporting scale for the IELTS General Training module in 1995; I only had the report by Charge and Taylor (1997) that this was done. Similarly, the change in the TOEFL Reading test where the vocabulary items were embedded in the reading passages seemed to be motivated by audience demand. Studies that were published about the change investigated the measurement properties of the old and new tests but did not discuss the changes in the test construct. Studies or discussions may have been conducted among the test developers, but I was unable to analyse them because they were not published. The reasons for the decisions on what discussions are published can vary, and some of them will be discussed in the next section.

The problem with published and unpublished development material can only be solved by analysing more varied material. Spolsky (1995) was able to show the influence of institutional forces through archival analyses of meetings and retrospective interviews of members who had been present. His analysis was thorough and the report constituted a book. O'Loughlin (1997) was able to analyse an item editorial committee's bases for test development decisions through participant observation. Because the focus of his study was comparability of tape-mediated and face-to-face testing, he only discussed this aspect of the item editorial committee's deliberations, but his report showed that this was a good potential source of data for practices in test development. Such data are likely to be be so rich that even a within-examination analysis and report would constitute a large single-case study. Nevertheless, the perspectives into test development that such studies can provide certainly complement the picture I was able to provide in the present study, because the object of analysis then is actual practice.

Another advantage of ethnographic data on test development is the increase in voices that are heard on the activities.

However, some caution is also due because the analysis and reporting of ethnographic data on tests can become complex, especially in a commercially and politically laden world such as the development of commercial examinations. Observation and interviews produce data on actual activities, and these can be contrasted with the principles that the boards claim to follow. Once the researcher has found a structure for presenting the data and a stance from which the report is written – not a small challenge – the report can make a substantial contribution to existing research. Politically, however, it may not be in the testing boards' interests that such analyses are published, and this may give rise to conflicts between the board and the researcher and/or the board and the employees who provide the data. All can be seen as "owners" of the data, and the complex of loyalties and interests may result in difficulties for reporting. At the centre of these is the notion of criticism, for which there is plenty of potential in the data. Criticism is difficult socially, and while it may be focused on substantive issues in a research report, it is easily interpreted as criticism of people and institutions. In addition expertise in language testing, applied linguistics, and psychometrics, such a report would probably call for expertise in social psychology. This is not to say that such research should not be pursued; on the contrary, there are clearly a whole range of contributions to be made here. The preparation for ethnographic analysis of test development must be thorough, however. Observation and interviews can also complement analyses such as the one conducted in the present study. However, in addition to the political problems discussed above, it would have been difficult to make the data comparable because of the differences in the time of the cases. Participant views gathered now on developments in the late 1980s (TOEFL Reading) are likely to contain much more rationalisation than those on ongoing developments (TOEFL 2000). The upshot for future research is that participant views must be gathered while the development is going on.

9.5.2 Theoretical development and testing traditions

The nature of theoretically acceptable practice changes over time. An illustration of this was the discussion of developments in validity theory in Chapter 3 above. In the case of test development theory, Spolsky (1977, 1995) has distinguished three different phases: the traditional or pre-scientific, the modern or psychometric-structuralist, and the postmodern or psycholinguistic-sociolinguistic. The phases are related to theories of

language ability, types of items presented, the criteria employed to assess the quality of the test. The traditional period is associated with extended answers, expert judgement to assess them, and little investigation of reliability or comparability of judgements. The psychometric period is associated with objectively scored discrete point items, while the postmodern period is associated with integrated testing and detailed, contextualised information on the construct. The overall pattern in the latter half of the twentieth century was for language tests to move away from the test methods associated with the modern phase and to place increasing value on the post-modern ways of defining the construct to be measured. Some of the differences between the tests studied above may be due to the different stages of theoretical development when they were first developed.

This argument is based on two assumptions: firstly, the stage of theoretical development matters when the format and development rationale for a test are being considered, and secondly, the critical stage when it matters is initial test development. Support for both assumptions is found in the case analyses. One of the tasks of the developers of IELTS was to revise the test's "outdated" construct. The test developers wanted the test to be up-to-date. Similarly, one of the arguments used by the developers of TOEFL 2000 was that they consulted current theories of language learning and communicative competence, because they believed that it was important that the test was up-to-date. Both arguments were made at the phase of initial test development, and no comparable arguments were published on the operational test forms.

The desire to be up-to-date builds on a belief that the current state of development is an improvement over the past. However, the argument is also practical. Assuming that fields such as language education hold the same belief and evolve accordingly, timely revisions to examinations help them fit well with the educational and other social practices around them.

The wish to be up-to-date, associated with constant if slow development of theory, means that examinations go through periodic revisions. Alderson (1986:96) has proposed that the life cycle of a language test might be 12-15 years. Although the period was longer, one such development was clear in the present thesis, since I analysed the TOEFL test at two different stages of development. The view of the construct assessed in the two versions is different, and this can partly be explained by the stages of theoretical development.

Acceptable practice can also vary across cultures, and in the case of language testing, one of the central distinctions that has been discussed at least in the literature published in English is the difference between UK and

US traditions (eg. Alderson 1987, Bachman et al. 1995). Differences are found in the roles and functions of testing agencies, test development procedures, and scoring and score interpretation. British examination boards tend to have strong links with educational programs and educational policy-making and thus provide certificates that indicate achievement in learning, whereas US testing agencies largely work independently of educational policymakers with a role of providing independent, reliable, and objective information for decisions that require evaluation. The political difference in the roles leads to differences in test development practices (see Bachman et al. 1995:15-18). In the United Kingdom, test development tends to be based on expert judgement whereas in the United States, decisions on examination quality are typically made on the basis of statistical analysis of pretest data. In the United Kingdom, assessment relies on the expertise of assessors, while in the United States, statistical reliability is used as the quality criterion in assessment practice. Alderson et al.'s (1995:256-259) evaluation of examination board practices in the United Kingdom in the mid 1990s indicated that the trend was still evident. It is possible that the assessment culture directs publication practice. Thus, it is possible that psychometric evaluations were conducted on the British-Australian based IELTS, but the results were not published, because the examination developers did not feel the need for it.

While historical development and cultural adherence may explain why there were differences between the cases investigated above, it does not annul the fact that there were differences. Moreover, my focus on certain types of difference is probably motivated by the current assessment context, and from this vantage point, it is possible to evaluate the usefulness of existing practice whatever its historical or cultural basis for the aims that assessment specialists currently consider important. The current, post-modern stage of test development and validation theory promotes communicative competence and the importance of detailed construct information. According to current beliefs, this is partly motivated by a desire to give detailed construct-related feedback to examinees. While I recognise that the tests I investigated came from different traditions, my analysis did not focus on the stages of development but on current practices.

9.5.3 Test development brief: resources, conditions, and constraints

Tests are developed for different purposes and within different political and practical contexts. The purposes of TOEFL and IELTS are quite similar, but the political and practical contexts for their development undoubtedly

influenced their format. Some dimensions of this situation were touched on above. In the discussion of test development theory in Chapter 3, I took up some of the concerns under conditions and constraints, and I also defined them in each of the case reports. Such considerations influenced possible test formats, score reports, and scoring practices.

Consideration of the size of the examinee population led the developers of the TOEFL test to decide on selected response tasks. Similar considerations led the developers of IELTS to decide on selected response and short, clerically markable, constructed response items. The difference may partly be explained by population size, but partly also by examination tradition. The British board may have found it difficult to accept a single selected-response test type only. Possibly for reasons related to assessment culture, they also found it impossible to exclude tests of writing and speaking, while practical considerations of cost led them to require single marking. The Speaking team for TOEFL 2000 suggested that the size of the examinee population required that the test should be semi-direct and possibly machine-assessed. Most if not all of these practical considerations led to differences in the construct definitions for the tests, and thus influenced my results. In fact, they formed part of my results, but they were not a central concern. The current values in educational measurement and assessment support some aspects of both traditions, namely the performance orientation of the British tradition and the checking of measurement quality that forms the core of the American assessment tradition. In fact, current measurement theory advocates the combination of both, but it remains to be seen whether any of the tests I investigated will implement this in the future.

9.6 Summary

Each of the case studies was concerned with a unique test development project, and each report revealed a complex network of activities performed by the test developers to build quality in their test. The individual case analyses were summarized at the end of each case report, and the cross-case comparisons were discussed above. In this section, I will present a brief summary of the main issues raised both within and across the cases. As was pointed out above, the results concern reported practice.

The case reports revealed differences in what "quality" meant for each group of test developers. This followed the categories that I had used for selecting the cases. That is, for TOEFL Reading, quality was primarily psychometric, for IELTS it was primarily theoretical or utilitarian, and for

TOEFL 2000, all these concerns were researched and the results were observed when test development decisions were made. The particular emphasis in each case led to differences in the main kind of information that could be made available on the basis of the scores. In the case of TOEFL Reading, this was mainly measurement information, a numerical value on the TOEFL score scale. The development practices supported the replicability of the assessment; parallel tests would give reported scores as similar as possible. In the case of IELTS, the premium was placed on the appropriacy of the test in terms of what was to be assessed and the usefulness of the information for score users. Although measurement quality is implied when scores are used in decision-making, the published reports did not concentrate on the measurement quality of the test. In the case of TOEFL 2000, the scores were to deliver detailed *and* reliable information on the abilities of the examinees as evidenced in their test performances. This was supposed to be useful both for selection purposes and for placement and diagnosis.

The differences in emphasis, especially between TOEFL Reading and IELTS, did not mean exclusive attention to one type of data rather than another in either case. Theoretically based construct information was used in both cases to guide test construction, while in validation, both systems used quantitative analyses. However, in the TOEFL Reading case, test construction procedures combined theoretical construct information with psychometric indicators, and in the IELTS case, validation studies, especially from the early stages of test development, included theoretical considerations of the construct. In both cases, the publicity material for the test made use of the published parts of the verbal construct definition. This meant that the construct was defined briefly in the TOEFL case and in a more detailed fashion for IELTS.

In TOEFL 2000, ways and means were specifically sought to combine theoretical and measurement perspectives in the definition of the construct. Although the objects studied included total and section scores, as in TOEFL, and the analysis of theory, as in IELTS, the main effort concentrated on the analysis of tasks. Tasks were analysed as communicative environments, on the one hand, and as representatives of points or regions on a measurement scale, on the other. The contextual, discourse, and performance requirement properties of the items were analysed to develop connections between indices of task content properties and empirical item difficulty information. A trial with the Reading section proved quite successful, while the report from the working group on

Speaking indicated that more work was required to make a similar approach work for this skill.

In terms of development and validation activities, the cases were largely shown to follow the expectations of the model of test development and validation that I presented in the case study framework. The activities of test development and validation were closely connected before the publication of the test and became more independent after it. In terms of the role of the verbal construct definition, the analysis revealed that it was more closely reflected in validation practices than in test development because the construct definitions provided the basis for content categories of items in all the cases reported. Validation, in contrast, was score-focused in TOEFL Reading and more construct-focused in IELTS. In TOEFL 2000, there were implications that both theoretical and measurement indicators would be used in both test development and validation. Analysed in this way, the construct definition was a relevant consideration in all the three cases, but the way in which it was expressed in the validation studies depended on the way in which the construct was viewed, as measurement-based, theoretical construct-based, or both.

Part Three
Concluding discussion

10 CONCLUDING DISCUSSION

The aim of this study is to clarify the role of construct definition in language test development and validation. From personal work experience and from reading theoretical texts, I had the impression that the three topics – test development, validation, and construct definition – were closely related but that they were not always treated together as a coherent whole. In my work as a test developer I had found that construct definition was difficult and that, theoretical recommendations aside, to present the early products of a validation process as validity evidence was challenging and guidelines for how to do it well were either not detailed or not easy to understand. I wanted to clarify how test development and validation could and should be done.

In the two parts of the present thesis, I addressed the combination of test development, validation, and construct definition from two perspectives, that of recommendations from theory and that of reports of practice. In Part One I treated the topic in three theme-specific chapters because this was the best way that I could find to analyse recommendations and alternatives from a test developer perspective, given that the texts about these topics were not always written from this perspective. In Part Two of the thesis I used the findings from Part One to build a framework of analysis for test development and validation practice. I applied the framework to three cases for which published reports on development and validation were available. The cases differed in terms of their approach to construct definition. The results of the case analysis were summarised at the end of each chapter and the cross-case analyses were discussed in the previous chapter. In this chapter, having analysed reports of practice, I return to the questions and findings of Part One of the thesis in terms of test development, validation, and construct definition. Finally, I will discuss the limitations of the study and present suggestions and directions for future research.

10.1 Recommendations for test development revisited

The analysis of recommendations for test development from theory in Chapter 2 concentrated on the stages of test development, the qualities to be observed in development, and the instructions for validation. The results indicated that there was a broad consensus on the stages of test development and that the qualities to be observed in the work were

reliability, validity and practicality. The advice for validation was that it should be implemented as a process alongside test development, that validation was concerned with the theoretical definition of the constructs assessed in the test and, after the test is published, with the empirical properties of test scores and their relationships with other scores and ability indicators. The theoretical definition provided the scientific backing for the score meaning and guided questions to be asked in validation, while scores provided data for the investigations. The theorists considered test development and validation very closely related, and although they concentrated on development and treated validation in less detail, all authors emphasized that the validity of contextualised score interpretations was the most important quality criterion that could be applied to them. This made validation important for test development.

In the framework used to guide the analysis of practice above, I implemented the shared view of the stages of test development. A procedural view of test development and validation was presented in Figure 3 in Chapter 5 and will not be re-presented here, but I will briefly discuss the related findings from Part Two. The reports of practice were largely coherent with the theoretical model. Moreover, the reports of practice provided concrete examples of what it meant that the logical stages of test development were interconnected and iterative, especially in initial test development. The outcomes of all activities influenced each other, and the result was an improved draft test. The reports of practice also showed that the distinction between pre-publication and post-publication test development was relevant especially for revisions in the construct definition and the task specifications. Because scores from operational tests had to be comparable across administrations, the specifications, tasks, and rules for score reporting had to remain stable. Changes were made through formal revisions. According to the reports of practice, the start of formal revisions was a policy matter that was connected with perceptions of the acceptability and appropriacy of the test construct, not a direct result of scientific criticism. Nevertheless, when the revision was implemented, construct definition was addressed in a scientific sense with a desire to improve both the definition and its implementation.

In section 2.10.3 in Chapter 2, I summarised the principles recommended in theory for good practice in a set of desirable goals, and Part Two of the thesis provided some examples of practical test development activities that implement them. The goals and possible means for reaching them are summarised in Table 6.

Table 6. Goals for test development and means for reaching them	
Goals	**Means**
to measure the right thing	- define skills to be assessed in detail - define task characteristics and task rubrics - check acceptability and appropriacy through peer and test policy board comments - analyse tasks from the perspective of task demands to make closer description of skills - refine tasks through peer comments - use empirical information from trialling to select best tasks - use empirical information from trialling as criterion when test forms are constructed
to measure consistently	- use empirical item information from trialling to select best tasks - check that all new test forms follow content and statistical criteria - monitor standardisation of administration including the administration of interactive speaking tests - monitor standardisation of rating when human rating is used - monitor measurement properties of actual tests and make revisions in methods of construction and/or analysis as necessary
to measure economically	- analyse possible overlap through eg. factor analysis - remove all overlapping test sections that you can provided that you can deliver the scores that users need and provided that measurement properties do not suffer - fit as many items in test time as possible but monitor speededness
to provide comparable scores across administrations	- follow standardised administration procedures - monitor reliability - use well-documented methods for score conversion and test form equation
to provide positive impact and avoid negative consequences	- predict possible consequences and analyse realised consequences - ensure that negative consequences cannot be traced to test invalidity - consult and observe learners, teachers, materials writers, curriculum designers and researchers as sources of data on possible washback
to provide accountable professional service	- document all procedures carefully - provide advice for score interpretation - report measurement properties of reported scores

The reports of practice indicated that not all principles were equally important in all the cases investigated. It is likely that the observation or non-observation of the principles was guided by the values of the test developers or test development boards. In the case of consequences of measurement use, it is possible that the absence of empirical studies was related to the complexity of the issues, the ensuing complexity and labour-intensiveness of possible empirical designs, and the fact that theory recommends that responsibility for such studies is shared between test developers and test users, which leaves the responsibility for such studies open. Overall, in a theoretical sense, all the principles are highly desirable and they are also included in current professional standards, for instance the *Standards for Educational and Psychological Measurement* (AERA 1999). However, these cannot usually be enforced. This leads to a situation where the score users must detect when principles are not followed, and if there is choice, they can choose a test that delivers the kind of information that they need and that provides the quality they want. Thus, if a test reports numerical scores without an explanation of their meaning, the potential participant or user may be happy with it. If a test reports numerical scores and explanations of their meaning but no statistical information about the measurement quality of the test, another user may be satisfied with it. On the one hand, this places high demands on the sophistication of score users. On the other, it gives them some power; the test revisions reported in Part Two of the thesis were motivated partly by consumer demand. Professionally valid arguments for the quality of tests can be developed with reference to professional standards if the developers are committed to them. The current standards require both theoretical and psychometric information about the test.

In addition to documentation of development activities and score analyses, a very important document related to a test is its specifications. According to the development reports, this document is often internal to the testing board, at least after the test is published, but from the perspective of those who work on a test, it is a very important means in quality development for the test. The frameworks of test development discussed in Chapter 2 required that the specifications should contain the theoretical construct definition for the test and detailed guidelines for how it should be operationalized in test tasks and assessment criteria. The specifications are used in the internal evaluation of the test and in system development. Detailed records of their use were not analysed in Part Two because of the confidentiality of the document. However, the role of construct definition was discussed in the reports.

According to both advice from theory and reports of practice, the role of construct definition in test development is to guide the activities. It focuses the task developers' minds on "the right thing" and defines criteria for the range of aspects that must be covered in a test form. Observation of theoretical construct definition at all stages of test construction supports the creation of coherence between test tasks and assessment criteria. It also provides the theoretical basis for the test and, at least in principle, enables the test developers to develop theory-based hypotheses about how scores ought to vary under different examination conditions or between different groups. If designs like this are developed on the basis of known differences in the language learning background, these theoretically motivated studies could contribute to existing theoretical knowledge of the nature of language ability and possibly also language acquisition. Such a contribution would require that explicit information about the nature of the ability assessed in the test were available and that the hypotheses were developed on the basis of the variables identified in the construct definition.

10.2 Procedural view of validation

In Chapter 3 I set out to define what validation was, what test developers should validate, how the aim could be turned into a sequence of activities, and what the role of construct definition was in the process. I found that the definition of validation had evolved from its 1950s form and that the validity of tests, or rather test*ing*, was a complex property with several complementary approaches to its characterisation. At its core, validity denoted the meaning of the measure, but because the "measure" is a complex of tasks and criteria that have to be administered to derive scores for individuals, and because the current interpretation of validation includes the use of the scores, there are a large range of factors that can and do contribute to the meaning of the measure. According to the current professional standards, the aim of validation is to provide "a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use" (AERA 1999:9). This defines a very complex and challenging arena of activity.

Validation *is* challenging, but with a procedural view to its implementation, it is also something concrete and doable. The procedural view enables test developers to begin validation at the same time that test development begins, which is what theory advises that they should do. The reports of practice in Part Two of the thesis showed that this was in fact done, although not all stages or products of the work were always

published. Validation as a process involves careful consideration of test purpose and the intended context of use and thus situation-accommodated planning. At the early stages of test development, the planning concerns the nature of the skill that is to be assessed and the nature of the instrument through which the assessment will be made. For these definitions, the case reports showed that the developers used both existing theory and the developers' experience of the development process to help define the dimensions. The definitions were written in test frameworks and test specifications and realised through the properties of test tasks and assessment criteria.

Another strand in validation is the social dimension of score interpretation and score use. This side of validation addresses the fact that tests are used in society to give information and make decisions. According to validity theory, meanings on this side of validity investigations are seen as value-laden, and the responsibility of test developers and score users is to take account of the consequences of score use. In the reports of practice analysed in Part Two, one study (Bailey 1999) made proposals for the study of test impact when TOEFL 2000 would eventually be used, while none of the studies addressed the realised consequences of score use. Given that the studies were related to the development of the test and were mostly published by the test developers or examination boards, this is quite logical. Possible impact is something that boards can prepare for. Studies of realised score use are joint ventures between the test developers and score users and some might be conducted by score users alone. They might include critical social analysis of testing practices in contrast with other possible justifications for fair decision making.

Once test development is so far advanced that scores become available, these become a major type of data to be analysed in validation studies. Validity theory proposes a very wide range of methods to study the numerical properties of the scores and statistical backing for score generalization, on the one hand, and theoretically motivated studies of score relationships and score changes, on the other. These studies clarify the meaning of the scores and the stability of the meaning across populations and administration conditions. This is done by making theoretically motivated predictions of how scores should change in different conditions and then checking if they did, or analysing scores for sub-groups of examinees to see if there is desirable or undesirable group-specific variation. Among the studies I analysed in Part Two were some that compared scores for examinees from different linguistic and cultural backgrounds, for example, and some that analysed relationships between subtest scores.

One particular kind of study that seemed to lead test developers forward in the analysis of the constructs assessed involved a design where indicators of task characteristics were combined with indicators of task difficulty. Particularly the analysis of task demands, ie. what the examinees were required to do to perform successfully on the task seemed useful in the context of the TOEFL 2000 analyses of reading tasks. At the level of the whole test or a subtest, the next step would be the possibility to develop score reporting scales based on item properties. This possibility was discussed in Chapter 4 in the context of the Australian examples for reading and listening tests, and in Chapter 8 for TOEFL 2000. The development of such scales would require that items characterised by certain content properties would really cluster at different regions on a scale of item difficulty. If such a set of item properties were found, the properties could be used to describe scale levels in score reporting. The trials in TOEFL 2000 were not yet far enough advanced to show whether this was possible in practice. If it were, a detailed construct definition through the description of task properties would form the link between the nature of the skills assessed and the measurement properties of the test.

10.3 Construct definition

In the course of the present thesis, I have emphasized the properties and roles of the theoretical construct definition in test development and validation. I discovered that with the rise of construct validation in validity theory, theoretical construct definition had been promoted to a coequal status with the measurement properties of the test. It does not replace psychometric considerations, however. Since tests produce scores and scores should indicate the degree of examinee abilities in the skills tested, the theoretical definition is very closely linked to the psychometric properties of the scores. Current theory advises that both must be investigated in the test development and validation process and that neither can be omitted.

The questions that I posed in Chapter 4 about construct definition focused on the nature of the constructs that were identified in language testing theory and the range of alternatives that were available for language testers to describe the constructs related to their test. The results indicated that most if not all the approaches discussed could be categorised as interactionalist in orientation. Consistency in performance was seen to be explained by factors that described an individual's interaction with a context.

I discussed theoretical and empirical approaches to construct definition in Chapter 4 and contended that test developers could use both. The case reports in Part Two indicated that this is what the developers did. They considered a wide range of available theoretical literature and made an eclectic combination of several features from different texts. They also included consideration of stakeholder views and examinee needs, which I had not considered in Chapter 4. When they analysed the construct assessed in the draft or operational tasks, they used score data and task properties to define the construct in more detail. Detailed linguistic analysis of examinee performances or analysis of assessor perceptions were not used according to the case reports. This may have been because the cases were not representative of the whole range of practice in test development and validation, but also possibly because the results of such studies may be too complex for the comfort of examination developers who want to believe that their work is worthwhile.

When test developers worked on their test with the intent to clarify the construct assessed, the result was an improved assessment system, because tasks or criteria that needed to be changed were changed, and the detailed construct definition was refined in the process as well. However, this was not an obligatory part of the process. One of the three cases was a test where detailed theoretical definitions of the construct were not used in publicity material or in score reporting. This practice indicated the values of the test developers and their views about what was relevant score-related information. To spell out the implication more broadly, if the score is a single number with little if any verbal description of what it means, the numerical value of the score is considered the most important and meaningful aspect of the system. If detailed scores are reported in a score profile, some more information is considered relevant but the numerical values of scores are still the bearers of important information. The more verbal descriptors there are for the score meaning, the more overtly the test developers are committed to the theoretical definition of the construct assessed and its meaningfulness and usefulness to score users. It is important for them to report both the score and a description of the ability that it stands for. Similarly, if it is important for the test developers to report on the consistency and thus the dependability of the numerical scores, they will report reliability and standard errors of measurement.

The amount of score-related information that a test developer or testing board publishes is related to practical score use, on the one hand, and the social-political dimension of score use, on the other. The practical use of scores-as-numbers is usually decision-making. This can concern

large groups, and detailed verbal information about the content meaning of the scores might be cumbersome in such contexts. The practical uses of detailed verbal score reports are placement information for course providers and diagnostic information for teachers and participants. On the social-political dimension, the amount of information that is available about the test and scores is related to decisions about whether to use the test at all and how much confidence to place on the scores. With reports of reliability and standard errors of measurement, the test developers give score users information about the consistency and dependability of the scores. If this information is not provided, informed dependability judgements cannot be made; if this information *is* provided and it is favourable to the test, it is an argument for the quality of the test. With reports of the theoretical meaning of scores, the developers say what was assessed in the test and, by implication or possibly by outright statements, what was *not* assessed. Through this means they share power with score users as concerns the decision of whether to use the scores at all. Another type of decision that score explanation can inform is the judgement of what score levels might be critical or necessary for the purpose for which the decisions are made. If content information about score meaning is not provided, informed decisions about whether to use the test and what the critical levels are cannot be made on content grounds. If such information is provided, this can be used as an argument for the social accountability of the test and its developers. According to current standards in educational measurement, accountability requires that test developers provide information about reliability, and they should also be able to say what the scores from their test mean in terms of the abilities assessed. Through these means, the pain of choice is shared with score users.

In Chapter 4, two models were discussed that illustrate dimensions in the operationalization of constructs in test tasks. They were relevant for the analysis of the approaches to construct definition presented in the same chapter. Chapelle's (1998) model focused on the range of factors that must be defined to account for an interactionalist view of language ability, and the Kenyon-McNamara-Skehan model illustrated the factors in the testing and assessment process that can influence variations in scores. Chapelle's model was useful because it showed clearly the distinction between factors related to the learner and factors related to the context of language use. According to the interactionalist definition of language ability, both sets of factors influence performance consistency, which arises from the interaction of an individual with a context of language use. The context can include other individuals or physical and textual objects only. When an account of

performance consistencies is sought through test performance, both individual and contextual factors must be defined in the properties of the test, ie. test environment, test rubric, test input, expected response, and relationship between input and expected response. The Kenyon-McNamara-Skehan model of the testing and assessment process can be seen as an extension of the central, "test" part of Chapelle's model. It defines the dimensions in the test instrument when extended performances are assessed, especially in the assessment of spoken interaction.

Test tasks and assessment criteria define the context in which scores are produced. The statistical properties of the scores give numerical information about the score meaning in relation to other scores. To make the numbers meaningful, the interpretations must be connected with the abilities assessed. The case reports showed, as Chapelle's model predicted, that it may be possible to make this connection through task properties. As was discussed above, this may mean that score reporting scales can be defined in terms of task properties, if it is found that easy tasks can be described through some content characteristics and more difficult tasks through other characteristics. This concerns tests where task characteristics can be defined in detail in advance, such as reading or listening tests. Accordingly, the reports from the TOEFL 2000 project showed that progress was fastest in the reading test, where some analyses of task characteristics had been conducted and the percentage of task difficulty explained by a set of content characteristics was high, nearly 90%. In contrast, the speaking group was not nearly as far advanced in its analyses.

In performance assessment, the modelling of the variables that have a potential influence on scores is different in three senses. Firstly, the examinee performance is extended and the production of scores requires judgement of its quality. This distinguishes performance assessment from assessment methods oriented towards objective scoring. Secondly, within performance assessments, the definition of task requirements is different depending on the degree of task structuring. Thirdly, administration can vary where tests of speaking are concerned. At one extreme on this dimension are semi-direct tests of speaking where the input can be analysed in advance and the interaction during the test situation is not bidirectional. At the other are interactive assessment situations with two or more interactants and different configurations of participant relationships with respect to gender, power, or language ability. The two last dimensions interact and combine.

The fact that performance assessments require human judgement has several implications for dimensions to be accounted in score variation,

including different types of assessment scales and assessor perceptions of qualities that must be assessed in the performances. In Chapter 4 above, designs related to the study of such variation included the analysis of scores given for different tasks by different rater groups (Chalhoub-Deville 1995, 1997) and the analysis of examinee performances to produce effective scale descriptors to enhance comparability of scores given by different raters. The case studies did not include investigations of this aspect, but for tests where performance assessment is used, this would be an important area of study.

Task structuring can range from highly structured tasks, where the input and expected response can be defined fairly well before the test is administered, to open tasks, which allow negotiation and interpretation by examinees so that the task properties can vary between individuals. The degree of appropriateness or correctness of task interpretation may or may not be one of the performance features assessed; this has to be defined in the assessment criteria. In Chapter 4 above, Skehan's (1998) task characteristics were related to variation in task structuring. Skehan's task dimensions were referred to by the developers of the TOEFL 2000 Speaking framework (Butler et al. 2000), but no detailed designs for the analysis of their influence on performance or on item difficulty had been developed yet. Possible variation in task structuring also influenced the design decisions in the IELTS speaking test when the structured interview was chosen as the test method instead of an unstructured interview (Ingram and Wylie 1997).

The choice of whether a speaking test is administered in tape-mediated or face-to-face format and what types of face-to-face interaction it may include influence the degree to which the administration conditions and task properties can be defined in advance rather than in retrospect, which in turn has an effect on the degree to which the nature of the ability assessed can be defined without observing the test or transcripts of it. Research on the influence of interlocutor variation on examinee scores is only beginning to emerge. One study discussed in Part Two above, Brown and Hill (1998), was concerned with this aspect and contended that interlocutor style influenced the nature of the interaction in the test but scoring may mask some of the variation in test interaction, because raters seemed to compensate for interlocutor harshness. Given that comparable tests should be given to different test takers, this is what raters should do, although it would be a better argument for the quality of the test, if the developers could specify exactly how this is done. Other research in language testing on interlocutor variability is beginning to emerge, and this is a necessary area

of validation research for tests where the interactive mode is chosen for assessing speaking. Once there is some more information about the nature of interlocutor variability and its influence on test discourse, it will be possible for test developers to develop means to control for such variability and to train interlocutors and raters to avoid its effects.

On the basis of the case analyses, it seemed that the decision between tape-mediated and face-to-face modes for testing speaking was related to the values of the test developers, their views of the necessary aspects of the construct that had to be assessed, and practicality considerations. In the IELTS case, the decision was made to assess speaking in live interaction. The possibility for variation in task and assessment was controlled to the extent possible by the definition of task rubrics, by assessor training, and by the characteristics identified in assessment scales. Analyses indicated that administration did vary and the researchers recommended that it should be controlled better. The reliability of the speaking scores was not analysed or discussed. In TOEFL 2000, the developers' values seemed to direct them to the decision that the test would be indirect, but the decision had not yet been made when the newest report that I analysed was published.

The three perspectives into variation in tests and scores discussed above are important objects of study in the context of performance assessment. They are especially important from the construct perspective because of the relationship between constructs and performance consistency. Constructs are defined by this consistency, and in order to detect it in the first place and assess its meaningfulness for the intended object of measurement, the relevant properties of the assessment situation must be known to some extent. More research is needed into the relevant variations in this area.

The result of the focus on construct definition, and the value promoted by current validity theory, is that scores from tests should be reported with reference to the construct assessed. The question must be asked, however, of what purpose this serves. Who cares if constructs are not described and score interpretations are not described in terms of the abilities assessed? As discussed above, the rationale may be that users have a right to this information whether they choose to use it or not. The contrast is with the situation where scores are reported as plain numbers. In such a case, the score user is dependent on the procedures used by the testers to derive the scores and the cultural interpretation of the numerical scores from the test in question. Their relevance in the context where score use is considered must be assessed on the basis of the information available on the test and possibly on the basis of local validation studies, which may help

define local cut scores. However, without task or performance analysis, the plain scores will not give information about the conceptual basis on which the decisions are made, they are simply numbers "that are available". The importance of content information is a value argument, however. It can be confronted with the opposing argument that the construct information is not needed. The quality principle heeded when detailed information about the nature of the scores is provided is that more information is better. If some of it is ignored in decision making, the responsibility is the score user's. If some of the information is not provided, the test developers are also accountable.

10.4 Limitations of the present study

In the present study, principles and reported practice in language test development and validation have been considered. These concerns are on a "meta" level, ie. they are one step removed from actual professional practice. From personal experience I can say that actual test development work is more concrete and more complex. I analysed reported practice because this gave me a meta-level overview of practices in test development. I assumed that publication would mean that the developers regarded the activities reported as reasonably acceptable practice. Nevertheless, this level of analysis does not describe the day-to-day concerns of test development, and the results must be viewed in this light.

The data that I analysed in the present study was published research on test development, and some reference was also made to publicity material for the tests that were analysed. The range of data did not include policy level documents such as minutes of meetings, although Spolsky (1995) has shown that this type of data is important for explaining why some test development decisions were made and what considerations were taken into account when this was done. I chose my perspective because of my own interest and because of access to data, but the policy level would provide another interesting perspective. For the type of examination that I investigated, such a study would focus on national and international differences in educational policy. This was not the focus of the present study, but the findings must nevertheless be interpreted with the knowledge that the policy level has not informed the analysis.

Taking the study as it was, further limitations include its limited scope. With this I mean firstly that I only discussed three cases, one of which was a test section rather than a whole test. This does not represent the whole range of professional practice, but throughout the study I have

made careful note of this and I have not made generalizations to all test development. From another perspective, a limitation of the present study is that it represents one person's decisions on what material to include and one person's interpretation of the significant points in it, my own. I have tried to record carefully what material there is and state my reasons for why I discussed certain aspects of it. I defined my aim and focus in some detail in Chapter 1 and presented a set of research questions that I addressed. I stand by the discussions I presented and the conclusions I drew, but I note that these were guided by the questions I asked. The present study is one of ideas, and if someone else analysed the same material, they might raise different questions. This study concentrated on a topic that I considered important, and I have learnt more about it in the course of writing the thesis. It is hoped that its combination of theoretical issues and its suggestions for construct-related work in test development and validation can be of some use for other test developers as well, especially the model of test development and validation in figure 3 in Chapter 5 and the summary of goals of accountable measurement and means for addressing them in Table 6 above.

10.5 Directions for future research and practice

In the course of the present study, several topics have been raised for possible future research that would increase our understanding of the constructs assessed in language tests and the ways in which this understanding can be used in test development and validation. Some of the directions are more theoretical, some more related to concrete test development activities, and some combine both. All are related to the combination of a theoretical definition of the skills assessed and the expression of the abilities as numbers or score categories.

The most promising proposal for the linkage of the two dimensions of construct definition was the combination of an analysis of task properties and an analysis of task difficulty. Since this direction has been discussed several times above, a brief mention will suffice here, although the promises for construct description and score explanation are great. This approach was possible for highly structured tasks, such as typical reading or listening tests, where task properties and task demands can be analysed in detail. It requires that the right dimensions for task definition are found to explain task difficulty. For further development in score reporting, it requires that classes of task properties cluster at regions of the difficulty scale and that if the difficulty of some components is spread out, the spread can be

explained by additional task dimensions. Once some successful solutions for some sets of tasks are found, the generalizability of the construct dimensions identified can begin to be analysed. This area of research is appropriate both for researchers and for test developers, and an assessment of the usefulness of these analyses in different contexts would be useful for the development of the language testing as a field of research and practical activity.

A related area of research is studies that focus on the nature of performance assessment and the constructs assessed in it. The three dimensions of difference between selected response tasks and performance oriented tasks were use of human judgement, degree of task structuring, and variables in the administration of face-to-face speaking tests. Research in all of these areas has begun to appear, but more research is needed if we are to understand performance assessments with respect to the two dimensions that are important in construct definition: score variation and the nature of the abilities that the scores indicate. Current practice to develop quality in tests that use human judgement is to train raters, use at least two raters and average the scores given, and monitor their internal consistency and analyse and possibly compensate for their harshness. Ongoing research (eg. Lumley 2000, Tarnanen forthcoming) investigates the raters' reasons for the scores that they give. Comparable research is needed into the nature of the differences in skills assessed in tasks where extended performances are assessed in contrast to machine-scored tasks where responses are selected or very limited in scope. This research can arise from the two other dimensions, task structuring and variation in administration, and research in these has also begun to appear (eg. Berry 1997, Brown 1998, Foster and Skehan 1996). To be useful from the assessment point of view, such research should combine the two perspectives of construct understanding and score variation.

The nature of test-related validation activities was discussed to some extent in the present thesis, but mostly from the perspective of construct definition. Research on concrete cases in other avenues in validation is needed to assess the validator's tasks and determine the responsibilities of the test developer with respect to them. One area where I do not know of existing research of practical cases in language testing is Kane's proposal for building validity cases. Are such cases in fact built in language testing? By whom? Who is the audience? What factors make it necessary to conduct the studies, or what factors make it possible that such studies are not demanded? Such a study would constitute a practical evaluation of

validity theory from the perspective of examination publishing, and it might help define limitations in the responsibility of the test developer.

A related but more test development-oriented clarification of validation requirements would investigate the usefulness of the procedural view of validation for language test developers. Do testers know how to implement validation as a process? The case reports in the present study showed that at least some test developers do it; is this common? Are the guidelines clear? Do test developers get caught if they do not do it? Who cares, and are there any sanctions if validation is not done? I suspect that similarly to the evaluation of the results in the present study, the answers are related to the values of the test developers. Professional credibility might be one of the products of such work, and as was discussed at length above, concentration on the construct assessed improves the quality of the test. Some validation work on the nature of the construct is produced almost automatically in the initial development of tests when tasks and assessment criteria are written and revised. It might be possible to document such work even in retrospect if this is motivating for the test developer. The recasting of construct validation as careful planning and revision might be helpful in such contexts.

For the purposes of developing an understanding of the nature of test development work, an ethnographic approach to it would provide a complementary perspective to what has been provided in the present thesis and in earlier work. The amount and range of material potentially relevant for such a study, if it concentrated on an examination revision or the initial development of a new examination, would conceivably be so large that limitations would be required. The development of both IELTS and TOEFL 2000 took several years and the results were published in a series of articles and research reports, none of which included "true" ethnographic material. Perspectives such as Peirce's (1992, 1994) on the development of an individual reading passage and associated items are well defined and could be repeated. A similar design transferred to the process of initial test development might combine an ethnographic approach with a focus on the influence of policy and personal effort as shown in Spolsky's (1995) research. Such a policy perspective into the development of language tests might provide additional information about the nature of compromises made in the test development process and the way in which such compromises are made, and provide more data on the values that guide test development. As mentioned in Chapter 9, this type of research requires a systematic data gathering strategy that is probably best implemented while the development is going on rather than as a post hoc design. The stance developed through

a systematic design might help in meeting the possible political difficulties of reporting the results.

10.6 Conclusion

In the present thesis, an attempt has been made to clarify the nature of construct definition in test development and validation. The main outcome was the point that there had to be a combination of theoretical and numerical dimensions in construct definition to make it possible to explain score meaning. The purpose of language tests is to assess language ability. To interpret their scores, means must be available to connect each score or category with conceptual explanations of what they stand for. Progress towards this goal can be made through careful test development which combines an analysis of the measurement properties of the scores with an analysis of task demands in terms of the abilities assessed. Neither alone is sufficient: numerical information gives evidence of the quality of measurement and theoretical information gives evidence of the quality of test content. Both are needed for accountable language testing.

REFERENCES

ACTFL (American Council on the Teaching of Foreign Languages) 1986. *ACTFL proficiency guidelines*. American Council on the Teaching of Foreign Languages, Hastings-on-Hudson, NY.

AERA (American Educational Research Association), APA (the American Psychological Association), and NCME (the National Council on Measurement in Education 1999. *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Alderman, Donald L. and Paul W. Holland 1981. *Item Performance Across Native Language Groups on the Test of English as a Foreign Language*. TOEFL Research Reports 9. Princeton, NJ: Educational Testing Service.

Alderman, Donald L. 1981. *Language Proficiency as a Moderator Variable in Testing Academic Aptitude*. TOEFL Research Reports 10. Princeton, NJ: Educational Testing Service.

Alderson, J. Charles 1986. Innovations in language testing? In M. Portal (ed.) *Innovations in language testing*. Windsor, Berks.: NFER-Nelson, 93-105.

Alderson, J. Charles 1987. An overview of ESL/EFL testing in Britain. In J.C. Alderson, K.J. Krahnke and C.W. Stansfield (eds.) *Reviews of English Language Proficiency Tests*. Washington, D.C.: TESOL, 3-4.

Alderson, J. Charles 1988. New procedures for validating proficiency tests of ESP? Theory and practice. *Language Testing* 5(2), 220-232.

Alderson, J. Charles 1990. Testing reading comprehension skills (part two). Getting students to talk about taking a reading test (a pilot study). *Reading in a Foreign Language* 7 (1), 465-503.

Alderson, J. Charles 1991. Bands and Scores. In Alderson, J. C. and B. North (eds.) *Language testing in the 1990s*. London: Modern English Publications, McMillan. 71-86.

Alderson, J. Charles 1993. The relationship between grammar and reading in an English for academic purposes test battery. In Douglas and Chapelle (eds.) 1993, 203-219.

Alderson, J. Charles 1997. Models of language? Whose? What for? What use? BAAL 1996 Pit Corder memorial lecture. In A. Ryan and A Wray (eds.). *Evolving models of language*. British Studies in Applied Linguistics 12. Clevedon: Multilingual Matters LTD.

Alderson, J. Charles and Gary Buck 1993. Standards in Testing: A Study of the Practice of UK Examination Boards in EFL/ESL Testing, *Language Testing* 10(2), 1-26.

Alderson, J. Charles and Caroline Clapham 1992. Applied linguistics and language testing: A case study of the ELTS test, *Applied Linguistics* 13(2), 149-167.

Alderson, J Charles and Caroline Clapham (eds.) 1992. *Examining the IELTS Test: An account of the first stage of the ELTS revision project*. International English Language Testing System Research Report 2. Cambridge: The British Council, The University of Cambridge Local Examinations Syndicate amd the International Development Program of Australian Universities and Colleges.

Alderson, J. Charles and Caroline Clapham 1997. The General Modules: Grammar. In C. Clapham and C. Alderson (eds.), 1997, 30-48.

Alderson, J. Charles, Caroline Clapham, and Dianne Wall 1995. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J. Charles and Dianne Wall 1993. Does washback exist? *Applied Linguistics* 14, 115-129.

ALTE 1996. *Guide for Examiners. Draft 1*. User's guide to Council of Europe, 1996. Strasbourg: Council of Europe.

Anastasi, Anne 1954. *Psychological testing*. First edition. New York: McMillan.
Anastasi, Anne 1982. *Psychological testing*. Fifth edition. New York: McMillan.
Anastasi, Anne 1986. Evolving Concepts of Test Validation. *Annual Review of Psychology*, 37, 1-15.
Angelis, Paul J., Spencer S. Swinton, and William R. Cowell 1979. *The Performance of Nonnative Speakers of English on TOEFL and Verbal Aptitude Tests*. TOEFL Research Reports 3. Princeton, NJ: Educational Testing Service.
Angoff W.H. 1988. Validity: An Evolving Concept. In H. Wainer and H. I. Braun (eds.), 1988, 19-32.
Angoff, W.H. 1989. *Context Bias in the Test of English as a Foreign Language*. TOEFL Research Reports 29. Princeton, NJ: Educational Testing Service.
Bachman, Lyle 1986. The Test of English as a Foreign Language as a measure of communicative competence. In Stansfield, C (ed.) 1986, 69-88.
Bachman, Lyle 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
Bachman, Lyle 1991. What does language testing have to offer? *TESOL Quarterly* 25 (4), 671-704.
Bachman, Lyle, Fred Davidson, Kathryn Ryan, and Inn-Chull Choi 1995. An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study. Cambridge: CUP.
Bachman, Lyle, Antony Kunnan, Swathi Vanniarajan, and Brian Lynch 1988. Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing* 5 (2), 128-159.
Bachman, Lyle F. and Adrian S. Palmer 1982. The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16, 49-65.
Bachman, Lyle and Adrian Palmer 1996. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.
Bailey, Kathleen M. 1999. *Washback in Language Testing*. TOEFL Monograph Series 15. Princeton, NJ: Educational Testing Service.
Banerjee, Jayanti and Sari Luoma 1997. Qualitative Approaches to Test Validation. In Clapham, C. and D. Corson (eds.) 1997, 275-287.
Bardovi-Harlig, K. and T. Bofman 1989. Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition* 11, 17-34.
Berry, Vivien 1997. Gender and personality as factors of interlocutor variability in oral performance tests. Paper presented at the Language Testing Research Colloquium in Orlando, FL.
Bingham, W.V. 1937. *Aptitudes and aptitude testing*. New York: Harper.
Boldt, Robert F. 1988. *Latent Structure Analysis of the Test of English as a Foreign Language*. TOEFL Research Reports 28. Princeton, NJ: Educational Testing Service.
Boldt, Robert F. 1991. *Cross-Validation of a Proportional Item Response Curve Model*. TOEFL Technical Reports 4. Princeton, NJ: Educational Testing Service.
Boldt, Robert F. 1994. *Simulated Equating Using Several Item Response Curves*. TOEFL Technical Reports 8. Princeton, NJ: Educational Testing Service.
Boldt, Robert F. and Courtney 1997. *Survey of Standards for Foreign Student Applicants*. TOEFL Research Reports 57. Princeton, NJ: Educational Testing Service.
Boldt, Robert F. and Roy Freedle 1996.*Using a Neural Net to Predict Item Difficulty*. TOEFL Technical Reports 11. Princeton, NJ: Educational Testing Service.
Boldt, Robert F., D. Larsen-Freeman, M.S. Reed, and R.G. Courtney 1992. *Distributions of ACTFL Ratings by TOEFL Score Ranges*. TOEFL Research Reports 41. Princeton, NJ: Educational Testing Service.

Brown, Annie 1998. Interviewer style and candidate performance in the IELTS oral interview. Paper presented at the Language Testing Research Colloquium in Monterey, CA.

Brown, Annie and Kathryn Hill 1998. Interviewer Style and Candidate Performance in the IELTS Oral Interview. In Wood (ed.) 1998, 1-19.

Brown, James D. 1998. An investigation into appriaches to IELTS preparation, with particular focus on the academic writing component of the test. In Wood (ed.) 1998, 20-37.

Buck, Gary and Kikumi Tatsuoka 1998. Application of rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing* 15 (2), 119-157.

Butler, Frances A., Dan Eignor, Stan Jones, Tim McNamara, and Barbara K. Suomi 2000. *TOEFL 2000 Speaking Framework: A Working Paper*. TOEFL Monograph Series 20. Princeton, NJ: Educational Testing Service.

Carey, Patricia A. 1996. *A Review of Psychometric and Consequential Issues Related to Performance Assessment.* TOEFL Monograph Series 3. Princeton, NJ: Educational Testing Service.

Campbell and Fiske 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin* 56, 81-105.

Canale, Michael 1983. On some dimensions of language proficiency. In J. Oller (ed.) *Issues in Language Testing Research*. Rowley, Mass: Newbury House, 333-342.

Canale, Michael and Merrill Swain 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1 (1), 1-47.

Celestine, Cynthia and Cheah Su Ming 1999. The effect of background disciplines on IELTS scores. In Tulloh (ed.) 1999, 36-51.

Chalhoub-Deville, Micheline 1995. A contextualised approach to describing oral proficiency. Language Learning 45: 251-281.

Chalhoub-Deville, Micheline 1997. Theoretical models, assessment frameworks and test construction. Language Testing 14 (1): 3-22.

Chapelle, Carol A. 1994. Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, 10 (2), 157-187.

Chapelle, Carol A. 1998. Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (eds.) *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press, 32-70.

Chapelle, Carol A. 1999. Validity in language assessment. *Annual Review of Applied Linguistics* 19, 254-272.

Chapelle, Carol, William Grabe, and Margie Berns 1997. *Communicative Language Proficiency: Definition and Implications for TOEFL 2000*. TOEFL Monograph Series 10. Princeton, NJ: Educational Testing Service.

Charge, Nick and Lynda B.Taylor 1997. Recent Developments in IELTS. *English Language Teaching Journal* 51 (4), 374-380.

Chyn, Susan, K. Linda Tang, and Walter D. Way 1995. *Investigation of IRT-Based Assembly of the TOEFL Test.* TOEFL Technical Reports 9. Princeton, NJ: Educational Testing Service.

Clapham, Caroline 1993. Is ESP testing justified? In Douglas and Chapelle (eds.) 1993, 257-271.

Clapham, Caroline 1996a. *The development of IELTS. A study of the effect of background knowledge on reading comprehension*. Studies in language testing 4. Cambridge: Cambridge University Press.

Clapham, Caroline 1996b. What makes an ESP reading test appropriate for its candidates? In A. Cumming and R. Berwick (eds.) 1996, 171-193.
Clapham, Caroline 1997. The Academic Modules: Reading. In C. Clapham and C. Alderson (eds.) 1997, 49-68.
Clapham, Caroline. and J. C. Alderson 1997. Introduction. In C. Clapham and C. Alderson (eds.) 1997, 1-2.
Clapham, Caroline M. and J. Charles Alderson 1997. *Constructing and Trialling the IELTS Test. IELTS Research Report 3*. Cambridge: The British Council, University of Cambridge Local Examinations Syndicate, and International Development Program of Australian Universities and Colleges.
Clapham, Caroline M. and David Corson (eds.) 1997. *Language Testing and Assessment, Vol. 7 of the Encyclopedia of Language Education.* Dordrecht: Kluwer Academic Publishers.
Clark, John L.D. 1977. *The Performance of Native Speakers of English on the Test of English as a Foreign Language*. TOEFL Research Reports 1. Princeton, NJ: Educational Testing Service.
Cole, Nancy S. and Pamela A. Moss 1989. Bias in test use. In R.L. Linn (ed.) *Educational Measurement.* Third edition. New York: American Council on Education/Macmillan.
Coleman, Gayle and Stephen Heap 1998. The misinterpretation of directions for the questions in the academic reading and listening sub-tests of the IELTS test. In Wood (ed.) 1998, 38-71.
Cook and Campbell 1979. *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
Cotton, Fiona and Frank Conrow 1998. An investigation of the predictive validity of IELTS amongst a sample of international students studying at the University of Tasmania. In Wood (ed.) 1998, 72-115.
Council of Europe 1996. *Modern languages: Learning, Teaching, Assessment. A Common European Framework of reference. Draft 2 of a Framework Proposal.* Strasbourg: Council of Europe.
Council of Europe forthcoming. Common European Framework of Reference for languages: learning, teaching, assessment. Cambridge: Cambridge University Press.
Criper, Clive and Alan Davies 1988. *ELTS Validation Project Report, ELTS Research Report 1(i).* London and Cambridge: The British Council and University of Cambridge Local Examinations Syndicate.
Cronbach, Lee J. 1949. *Essentials of psychological testing.* New York: Harper.
Cronbach, Lee J. 1971. Test validation. In R.L. Thorndike (ed.) *Educational measurement*. Second edition. Washington, DC: American Council on Education, 443-507.
Cronbach, Lee J. 1975. Five decades of public controversy over mental testing. *American psychologist* 30, 1-14.
Cronbach, Lee J. 1980. Validity on parole: How can we go straight= New directions for testing and measurement: Measuring achievement over a decade. Proceedings of the 1979 ETS invitational conference. San Francisco: Jossey-Bass, 99-108.
Cronbach, Lee J. 1986. Social inquiry by and for earthlings. In D.W. Fiske and R. A. Schweder (eds.) *Metatheory in social science*. Chicago: University of Chicago Press, 83-107.
Cronbach, L. J. 1988. Five Perspectives on Validity Argument. In H. Wainer and H. Brown (eds.). 1988, 3-17.

Cronbach, Lee J. 1989. Construct validation after thirty years. In R. L. Linn (ed.) *Intelligence: Measurement theory and public policy*. Proceedings of a symposium in honor of Lloyd G. Humphreys. Urbana, IL: University of Illinois Press. 147-171.

Cronbach, Lee J. 1990. *Essentials of psychological testing*. Fifth edition. New York: Harper and Row.

Cronbach, Lee J. 1998. Commentary on Ernie House and Michael Scriven's presentations. In R. Davis (ed.) *Proceedings of the Stake symposium on educational evaluation*. Illinois: University of Illinois Press, 25-28.

Cronbach, Lee J. and P. E. Meehl 1955. Construct validity in psychological tests. *Psychological Bulletin* 52, 281-302.

Cumming, Alister 1996. Introduction: The concept of validation in language testing. In Cumming, A. and R. Berwick (eds.) 1996, 1- 14.

Cumming, Alister and Richard Berwick (eds.) 1996. *Validation in Language Testing*. Clevedon: Multilingual Matters.

Davidson, Fred and Brian Lynch 1993. Criterion-referenced language test development: a prolegomenon. In A. Huhta, K. Sajavaara and S. Takala (eds.), *Language testing: new openings*. Jyväskylä: Institute for Educational Research, 73-89.

Davies, Alan 1977. The construction of language tests. In J. P. B. Allen and A. Davies (eds.), *Testing and experimental methods. The Edinburgh course in applied linguistics*. Vol. 4. London: Oxford University Press, 38-194.

Davies, Alan 1990. *Principles of Language Testing*. Oxford: Blackwell.

Davies, Alan 1994. A case for plausible rival hypothese in language testing research. Paper given at Language Testing Research Colloquium, Washington DC, 5-7 March 1994.

Davies, Alan 1996. Outing the tester: Theoretical models and practical endeavours in language testing. In G. M. Blue and R. Mitchell (eds.) 1996. *Language and education*. British Studies in Applied Linguistics 11. Clevedon: Multilingual Matters, 60-69.

De Bot, Kees 1992. A bilingual production model: Levelt's "Speaking" model adapted. *Applied Linguistics*, 13, 1-24.

Douglas, Dan 1997. *Testing Speaking Ability in Academic Contexts: Theoretical Considerations*. TOEFL Monograph Series 8. Princeton, NJ: Educational Testing Service.

Duran, Richard P., Michael Canale, Joyce Penfield, Charles W. Stansfield, and Judith E. Liskin-Gasparro 1985. *TOEFL from a Communicative Viewpoint on Language Proficiency: A Working Paper*. TOEFL Research Reports 17. Princeton, NJ: Educational Testing Service.

Edgeworth, F. Y. 1888. The statistics of examinations. *Journal of the Royal Statistical Society* 53: 644-663.

Enright, Mary K., William Grabe, Keiko Koda, Peter Mosenthal, Patricia Mulcahy-Ernt, and Mary Schedl 2000. *TOEFL 2000 Reading Framework: A Working Paper*. TOEFL Monograph Series 17. Princeton, NJ: Educational Testing Service.

ETS 1997. *TOEFL Test and Score Manual*. Princeton, NJ: Educational Testing Service.

ETS 1998. *Computer-Based TOEFL Score User Guide*. 1998-1999 Edition. Princeton, NJ: Educational Testing Service. Available from http://www.toefl.org/dloadlib.html. Downloaded November 1, 1999.

ETS 1999a. *TOEFL 1999-2000 Information Bulletin for Supplemental TOEFL Administrations*. Princeton, NJ: Educational Testing Service.

ETS 1999b. *TOEFL 1999-2000 Information Bulletin for Computer-Based Testing*. Princeton, NJ: Educational Testing Service.

ETS 1999c. *The Researcher*, Spring 1999. Test of English as a Foreign Language. Princeton, NJ: Educational Testing Service.

ETS 2000a. *TOEFL 2000-2001 Information Bulletin for Supplemental TOEFL Administrations*. Princeton, NJ: Educational Testing Service.

ETS 2000b. *TOEFL 2000-2001 Information Bulletin for Computer-Based Testing*. Princeton, NJ: Educational Testing Service.

ETS 2000c. *TOEFL 2000 Website*. http://www.toefl.org/toefl2000/index.html. Downloaded November 1, 1999.

Frederiksen, J.R., and Collins, A. 1989. A systems approach to educational testing. Educational Researcher, 18 (9), 27-32.

Foster and Skehan 1996. The influence of planning and task type on second language performance. *Studies in Second Language Acquisition* 18, 299-323.

Foster and Skehan 1997. Modifying the task: the effects of surprise, time and planning type on task based foreign language instruction. *Thames Valley working papers in English language teaching*. Vol. 4.

Foulkes, J. 1997. The General Modules: Listening. In C. Clapham and J.C. Alderson (eds.) 1997, 3-13.

Fox, Janna, Tim Pychyl, and Bruno Zumbo 1993. Psychometric properties of the CAEL Assessment, I: An overview of development, format, and scoring procedures. In J. Fox (ed.), *Carleton papers in applied language studies*, Volume X. Ottawa: Carleton University.

Freedle, Roy and Irene Kostin 1993a. The prediction of TOEFL reading item difficulty: implications for construct validity. *Language testing* 10, 133-170.

Freedle, Roy and Irene Kostin 1993b. *The Prediction of TOEFL Reading Comprehension Item Difficulty for Expository Prose Passages for Three Item Types: Main Idea, Inference, and Supporting Idea Items*. TOEFL Research Reports 44. Princeton, NJ: Educational Testing Service.

Fulcher, Glenn 1996a. Testing tasks: issues in task design and the group oral. *Language Testing* 13 (1), 23-51.

Fulcher, Glenn 1996b. Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13, 208-238.

Garrett, H. E. 1937. *Statistics in psychology and education*. New York: Longmans, Green.

Ginther, April and Leslie Grant 1996. *A Review of the Academic Needs of Native English-Speaking College Students in the United States*. TOEFL Monograph Series 1. Princeton, NJ: Educational Testing Service.

Griffin, Patrick and Shelley Gillis 1997. Results of the trials: a cross national investigation. In C. Clapham and J.C. Alderson (eds.) 1997, 109-124.

Guilford, J. P. 1946. New standards for educational and psychological measurement. *Educational and psychological measurement*, 6, 427-438.

Guion, R. M. 1980. On trinitarian doctrines of validity. *Professional Psychology* 11, 385-396.

Hale, Gordon A. 1988. *The Interaction of Student Major-Field Group and Text Content in TOEFL Reading Comprehension*. TOEFL Research Reports 25. Princeton, NJ: Educational Testing Service.

Hale, Gordon A., Donald A. Rock, and Thomas Jirele. 1989. *Confirmatory Factor Analysis of the Test of English as a Foreign Language*. TOEFL Research Reports 32. Princeton, NJ: Educational Testing Service.

Hale, Gordon A., Charles W. Stansfield and Richard P. Duran 1984. *Summaries of Studies Involving the Test of English as a Foreign Language, 1963-1982*. TOEFL Research Reports 16. Princeton, NJ: Educational Testing Service.

Hale, Gordon A., Charles W. Stansfield, Donald A. Rock, Marilyn M. Hicks, Frances A. Butler, and John W. Oller, Jr. 1988. *Multiple-Choice Cloze Items and the Test of*

English as a Foreign Language. TOEFL Research Reports 26. Princeton, NJ: Educational Testing Service.
Halliday, M. A. K. and R. Hasan 1989. *Language, context, and text: Apects of language in a social-semiotic perspective*. Oxford: Oxford University Press.
Hamel, Jaques, Stéphane Dufour, and Dominic Fortin 1993. *Case study methods*. Qualitative Research Methods Volume 32. Newbury Park: Sage Publications.
Hamp-Lyons, Liz and Caroline Clapham 1997. The Academic Modules: Writing. In C. Clapham and J.C. Alderson (eds.) 1997, 69-80.
Hamp-Lyons, Liz and Barbara Kroll 1996. *TOEFL 2000 -- Writing: Composition, Community, and Assessment*. TOEFL Monograph Series 5. Princeton, NJ: Educational Testing Service.
Hasselgren, Angela 1998. *Smallwords and valid testing*. PhD thesis. Bergen: University of Bergen, Department of English.
Harris 1969. *Testing English as a second language*. New York: McGraw Hill.
Heaton, J. B. 1975. *Writing English language tests*. London: Longman.
Heaton, J. B. 1988. *Writing English language tests*. 2nd edition. London: Longman.
Henning, Grant 1987. *A guide to language testing*. Cambridge, Mass: Newbury House.
Henning, Grant 1988. An American view on ELTS. In A. Hughes, D. Porter and C. Weir (eds.) 1988, 85-92.
Henning, Grant 1991. *A Study of the Effects of Contextualization and Familiarization on Responses to the TOEFL Vocabulary Test Items*. TOEFL Research Reports 35. Princeton, NJ: Educational Testing Service.
Henning, Grant 1993. *Test-Retest Analyses of the Test of English as a Foreign Language*. TOEFL Research Reports 45. Princeton, NJ: Educational Testing Service.
Henning, Grant and Eduardo Cascallar 1992. A Preliminary Study of the Nature of Communicative Competence. TOEFL Research Reports 36. Princeton, NJ: Educational Testing Service.
Hicks, Marilyn M. 1989. *The TOEFL Computerized Placement Test: Adaptive Conventional Measurement*. TOEFL Research Reports 31. Princeton, NJ: Educational Testing Service.
Hill, Kathryn, Neomy Storch, and Brian Lynch 1999. A comparison of IELTS and TOEFL as predictors of academic success. In Tulloh (ed.) 1999, 52-63.
Hudson, Thom 1996. *Assessing Second Language Academic Reading From a Communicative Competence Perspective: Relevance for TOEFL 2000*. TOEFL Monograph Series 4. Princeton, NJ: Educational Testing Service.
Hughes 1989. *Testing for language teachers*. Cambridge: Cambridge University Press.
Hughes, Arthur, Don Porter and Cyril Weir (eds.) 1988. *ELTS validation project: Proceedings of a conference held to consider the ELTS validation project report*. Cambridge: The British Council and the University of Cambridge local Examinations Syndicate.
Huhta, Ari 1993. Teorioita kielitaidosta: Onko niistä hyötyä testaukselle? [Theories of language: are they useful for language testing?] In S. Takala (ed.), Suullinen kielitaito ja sen arviointi. Kasvatustieteiden tutkimuslaitoksen julkaisusarja B: Teoriaa ja käytäntöä 77. Jyväskylä: Kasvatustieteen tutkimuslaitos.
Hymes, Dell 1971. Competence and performance in linguistic theory. In R. Huxley and E. Ingram (eds.). *Language acuisition: Models and methods*. London: Academic Press, 3-24.
Hymes, Dell H. 1972. On Communicative Competence. In J. B. Pride and J. Holmes, *Sociolinguistics*. Penguin books, 269-293.
Ingram, David E. 1984. Report on the formal trialling of the Australian Second Language Proficiency Ratings (ASLPR). Australian Government Publishing Service, Canberra.

Ingram, David and Elaine Wylie 1993. Assessing speaking proficiency in the International English Language Testing System. In D. Douglas and C. Chapelle (eds.) 1993, 220-234.

Ingram, David and Elaine Wylie 1997. The General Modules: Speaking. In C. Clapham and C. Alderson (eds.) 1997, 14-29.

Jamieson, Joan, Stan Jones, Irwin Kirsch, Peter Mosenthal, and Carol Taylor 2000. *TOEFL 2000 Framework: A Working Paper*. TOEFL Monograph Series 16. Princeton, NJ: Educational Testing Service.

Jonson, Jessica L. and Barbara S. Plake 1998. A historical comparison of validity standards and validity practices. *Educational and psychological measurement* 58 (5), 736-753.

Kaftandjieva, Felly, Norman Verhelst and Sauli Takala 1999. *A manual for standard setting procedure*. Unpublished guideline document in the Dialang project. Jyväskylä: University of Jyväskylä.

Kane, M. 1992. An Argument-based Approach to Validity. *Psychological Bulletin*, 112 (3), 527-535.

Kenyon, Dorry 1992. Introductory remarks at a symposium called *Development and use of rating scales in language testing*. 14th Language Testing Research Colloquium, Vancouver, February 1992.

Kirsch, Irwin and A. Jungeblut 1992. *Profiling the literacy proficiencies of JTPA nad ES/UI populations: Final report to the department of labor*. Princeton, NJ: Educational Testing Service.

Lado, Robert 1961. *Language Testing*. New York: McGraw Hill.

Lazaraton, Anne 1992. The Structural Organization of a Language Interview: a Conversation Analytic Perspective, *System* 20, 373-386.

Lazaraton, Anne 1996. Interlocutor Support in Oral Proficiency Interviews: the Case of CASE, *Language Testing* 13(2), 151-172.

Lewkowicz, Jo 1997. Investigating authenticity in language testing. Unpublished PhD thesis, Department of Linguistics and the Modern English Language, Lancaster University.

Levelt, William 1989. *Speaking: from intention to articulation*. Cambridge, MA: MIT Press.

Linacre, J. Michael 1994. *FACETS. (Version 2.75)* [Computer software]. Chicago, IL: MESA Press.

Linn, Robert L. (ed.) 1989. *Educational Measurement*. New York: American Council on Education / McMillan.

Linn, Robert L. 1993. Educational assessment: Expanded expectations and challenges. *Educational evaluation and policy analysis* 15, 1-16.

Linn, Robert L. 1997. Evaluating the Validity of Assessments: The Consequences of Use, *Educational Measurement: Issues and Practice* 16(2), 14-16.

Linn, R. L., E.L. Baker, and S. B. Dunbar 1991. Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20 (8), 5-21.

Loevinger, 1957. Objective tests as instruments of psychological theory. *Psychological reports* 3, 635-694 (Monograph Supplement 9).

Lumley, Tom 2000. *The process of the assessment of writing performance: the rater's perspective*. Unpublished PhD thesis, Department of Linguistics and Applied Linguistics, University of Melbourne, August 2000.

Lumley, Tom and Tim McNamara 1995. Rater characteristics and rater bias: implications for training. Language Testing 12(1), 55-71.

Lynch, Brian and Fred Davidson 1994. Criterion-referenced language test development; linking curricula, teachers and tests. *TESOL Quarterly*, 28, 727-743.

Maguire, Thomas, John Hattie and Brian Haig 1994. Construct validity and achievement assessment. *The Alberta Journal of Educational Research* XL (2), 109-126.

Manning, Winton H. 1987. *Development of Cloze-Elide Tests of English as a Second Language*. April 1987. TOEFL Research Reports 23. Princeton, NJ: Educational Testing Service.

McDowell, Clare and Brent Merrylees 1998. Survey of receiving institutions' use and attitude to IELTS. In Wood (ed.) 1998, 116-139.

McKinley, Robert L. and Walter D. Way 1992. *The Feasibility of Modeling Secondary TOEFL Ability Dimensions Using Multidimensional IRT Models*. TOEFL Technical Reports 5. Princeton, NJ: Educational Testing Service.

McNamara, Tim 1995. LT/AppLing article.

McNamara, Tim 1996. *Measuring second language performance*. London: Longman.

Mehrens, W. A. 1997. The Consequences of Consequential Validity, *Educational Measurement: Issues and Practice* 16(2), 16-18.

Meirion, Beryl E. 1998. Rating oral proficiency tests: A triangulated study of rater thought processes or inside raters' heads. Poster presentation at 20th Language Testing Research Colloquium, Monterey, CA.

Merrylees, Brent ja Clare McDowell 1999. An Investigation of Speaking Test Reliability with Particular Reference to Examiner Attitude to the Speaking Test Format and Candidate/Examiner Discourse Produced. In Tulloh (ed.) 1999, 1-35.

Messick, S. 1975. The Standard Problem: Meaning and Values in Measurement and Evaluation. *Americal Psychologist*, 30, 955-966.

Messick, S. 1980, Test validity and the ethics of assessment. *American Psychologist*, 35, 1012-1027.

Messick, S. 1981a, Constructs and their Vicissitudes in Educational and Psychological Measurement. *Psychological Bulletin* 1981, 89 (3), 575-588.

Messick, S. 1981b. Evidence and Ethics in the Evaluation of Tests. *Educational Researcher* 10 (9), 9-20.

Messick, S. 1984, The psychology of educational measurement. *Journal of Educational Measurement* 21, 215-237.

Messick, S. 1988. The Once and Future Issues of Validity: Assessing the Meaning and Consequences of Measurement. In Wainer, H. and H. Braun (eds.) *Test Validity*. 33-45.

Messick, S. 1989a. Validity. In R. L. Linn (ed.) *Educational Measurement*. Third edition. New York: American Council on Education / McMillan. 13-103.

Messick, S. 1989b, Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, 18 (2), 5-11.

Messick, S. 1994. The Interplay of Evidence and Consequences in the Validation of Performance Assessments. Educational researcher, Vol. 23 No. 2, pp. 13-23.

Messick, S. 1995a. Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist*, 50 (9), 741-749.

Messick, Samuel 1995b. Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice* 14, 5-8.

Messick, S. 1996. Validity and washback in language testing. *Language Testing*, 13 (3), 241-256.

Millman, Jason and Jennifer Greene 1989. The specification and development of tests of achievement and ability. In R. L. Linn (ed.) *Educational Measurement*. Third edition. New York: American Council on Education / McMillan. 335-366.

Mok, Magdalena, Nick Parr, Tony Lee, and Elaine Wylie 1998. A comparative study of the IELTS and Access test results. In Wood (ed.) 1998, 140-165.

Moore, Tim and Janne Morton 1999. Authenticity in the IELTS academic module writing test: a comparative study of Task 2 items and university assignments. In Tulloh (ed.) 1999, 64-106.

Moss, Pamela 1992. Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment. *Review of Educational Research* 62(3), 229-258.

Moss, Pamela A. 1994. Can There Be Validity Without Reliability?, *Educational Researcher* 64, 5-12.

Moss, P. 1995. Themes and Variations in Validity Theory. *Educational Measurement: Issues and Practice* 14, 5-13.

Munby, John 1978. *Communicative syllabus design*. Cambridge: Cambridge University Press.

Nattinger, J. R. and J. S. DeCarrico 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nevo, Baruch 1985. Face validity revisited. *Journal of Educational Measurement* 22 (4), 287-293.

Oller, John W. jr 1979. Language tests at school. London: Longman.

Oller, John W. jr 1986. Communication theory and testing: what and how. In Stansfield, C. (ed.) 1986, 104-155.

O'Loughlin, Kieran 1995. Lexical Density in Candidate Output, *Language Testing* 12(2), 217-237.

O'Loughlin, Kieran 1997. *The Comparability of Direct and Semi-Direct Speaking Tests: A Case Study*. Unpublished PhD thesis, Department of Linguistics and Applied Linguistics, University of Melbourne.

Oltman, Philip K. and Lawrence J. Stricker 1991. *Developing Homogeneous Scales by Multidimensional Scaling*. TOEFL Technical Reports 1. Princeton, NJ: Educational Testing Service.

Oltman, Philip K., Lawrence J. Stricker, and Thomas Barrows 1988. *Native Language, English Proficiency, and the Structure of the Test of English as a Foreign Language*. TOEFL Research Reports 27. Princeton, NJ: Educational Testing Service.

Pavlou, Pavlos 1995. *Assessing different speech interactions in an oral proficiency test*. Unpublished PhD thesis, Georgetown university.

Pawley, Andrew and Frances Hodgetts Syder 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. C. Richards and R. Schmidt (eds.) *Language and communication*. London: Longman. 191-226.

Peirce, Bonny Norton 1992. Demystifying the TOEFL Reading Test. *TESOL Quarterly* 26 (4), 665-691.

Peirce, Bonny Norton 1994. The Test of English as a Foreign Language: developing items for reading comprehension. In Hill, C. and K. Parry (eds.) 1994. *From Testing to Assessment: English as an International Language*. London: Longman, 39-60.

Pike, Lewis W. 1979. *An Evaluation of Alternative Item Formats for Testing English as a Foreign Language*. TOEFL Research Reports 2. Princeton, NJ: Educational Testing Service.

Popham, 1978. *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice Hall.

Popham, 1981. *Modern educational measurement*. Englewood Cliffs, NJ: Prentice Hall.

Popham, J. W. 1997. Consequential Validity: Right Concern—Wrong Concept, *Educational Measurement: Issues and Practice* 16(2), 9-13.

Popper, Karl 1968 (1959). *The logic of scientific discovery*. New York: Harper & Row.

Powers, Donald E. 1980.*The Relationship Between Scores on the Graduate Management Admission Test and the Test of English as a Foreign Language.* TOEFL Research Reports 5. Princeton, NJ: Educational Testing Service.

Robinson 1995. Attention, memory, and the noticing hypothesis. *Language Learning* 45 (2), 283-331.

Ross, Steven 1992. Accommodative Questions in Oral Proficiency Interviews, *Language Testing* 9, 173-186.

Rulon, P. J. 1946. On the validity of educational tests. *Harvard Educational Review* 16, 290-296.

Ryan, Kathryn and Lyle Bachman 1992. Differential item functioning on two tests of EFL proficiency. *Language Testing* 9 (1), 12-29.

Savignon, Sandra 1983. *Communicative competence: theory and classroom practice.* Reading, MA: Addison-Wesley.

Schedl, Mary, Ann Gordon, Patricia A. Carey, and Linda K. Tang 1996. An Analysis of the Dimensionality of TOEFL Reading Comprehension Items. TOEFL Research Reports 53. Princeton, NJ: Educational Testing Service.

Schedl, Mary, Neal Thomas, and Walter Way, 1995. An Investigation of Proposed Revisions to Section 3 of the TOEFL Test. TOEFL Research Reports 47. Princeton, NJ: Educational Testing Service.

Schmidt, Richard W. 1990. The role of consciousness in second language learning. *Applied Linguistics* 11 (2), 129-158.

Secolsky, C. 1989. *Accounting for Random Responding at the End of the Test in Assessing Speededness on the Test of English as a Foreign Language.* TOEFL Research Reports 30. Princeton, NJ: Educational Testing Service.

Shepard, Lorrie A. 1993. Evaluating Test Validity. *Review of Research in Education*, 19, 405-450.

Shepard, Lorrie A. 1997. The Centrality of Test Use and Consequences for Test Validity, *Educational Measurement: Issues and Practice* 16(2), 5-8; 13; 24.

Shohamy, Elana 1997. Critical Language Testing and Beyond. Plenary address given at the annual conference of the American Association of Applied Linguistics, Orlando, Florida, March 1997.

Skehan, Peter 1998a. Processing perspectives to second language development, instruction, performance, and assessment. *Working papers in applied linguistics*, Thames Valley University, London, Vol 4, 70-88.

Skehan, Peter 1998b. *A cognitive approach to language learning.* Oxford: Oxford University Press.

Skehan, Peter and Pauline Foster 1997. The influence of planning and post-task activities on accuracy and complexity in task-based learning. *Language Teaching Research* 1/3, 1997.

Skehan and Foster 1998. Task type and processing conditions as influences on foreign language performance. *Working papers in applied linguistics*, Thames Valley University, Londong, Vol 4, 139-188.

Spolsky, Bernard 1990. The Prehistory of TOEFL. *Language Testing*, 7 (1), 98-118.

Spolsky, Bernard 1995. *Measured Words. The development of objective language testing.* Oxford: Oxford University Press.

Stansfield, Charles W. (ed.) 1986. *Toward Communicative Competence Testing: Proceedings of the Second TOEFL Invitational Conference*. TOEFL Research Reports 21. Princeton, NJ: Educational Testing Service.

Swinton, Spencer S. 1983. *A Manual for Assessing Language Growth in Instructional Settings*. TOEFL Research Reports 14. Princeton, NJ: Educational Testing Service.

Swinton, Spencer S. and Donald E. Powers 1980. *Factor Analysis of the Test of English as a Foreign Language for Several Language Groups*. TOEFL Research Reports 6. Princeton, NJ: Educational Testing Service.

Tang, K. Linda 1996. *Polytomous Item Response Theory (IRT) Models and Their Applications in Large-Scale Testing Programs: Review of Literature.* TOEFL Monograph Series 2. Princeton, NJ: Educational Testing Service.

Tang, K. Linda and Daniel R. Eignor 1997. *Concurrent calibration of dichotomously and polytomously scored TOEFL items using IRT models.* TOEFL Technical Reports 13. Princeton, NJ: Educational Testing Service.

Tang, K. Linda, Walter D. Way, and Patricia A. Carey 1993. *The Effect of Small Calibration Sample Sizes on TOEFL IRT-Based Equating.* TOEFL Technical Reports 7. Princeton, NJ: Educational Testing Service.

Tarnanen, Mirja forthcoming. *Arvioija vaokiilassa: kirjoitelmien arviointi arvioijan näkökulmasta.* [The rater in focus: The assessment of writing from the rater's perspective.] PhD manuscript, Centre for Applied Language Studies, University of Jyväskylä, Finland.

Thorndike 1903. *Heredity, correlation and sex differences in school ability*. New York: Colombia University.

Toulmin, S., R. Rieke and A. Janik 1979. *An introduction to reasoning.* New York: Macmillan.

Tulloh, Robyn (ed.) 1999. *IELTS International English Language Testing System Research Reports 1999.* Volume 2. Canberra: IELTS Australia Pty Limited.

Turner, Carolyn and Jack Upshur 1996. Developing rating scales for the assessment of second language performance. In G. Wigglesworth and C. Elder (eds.) *The language testing cycle: from inception to washback.* Australian review of Applied Linguistics, Series S number 13. Melbourne: ARAL, 55-79.

UCLES 1996. *IELTS annual report 1995*. Cambridge: University of Cambridge Local Examinations Syndicate, the British Council, and IDP Education Australia: IELTS Australia.

UCLES 1999. *The IELTS Handbook 1999*. Cambridge: The University of Cambridge Local Examinations Syndicate, The British Council, and IDP Education Australia: IELTS Australia.

UCLES no date. *IELTS Annual Review 1997/8.* Cambridge: The University of Cambridge Local Examinations Syndicate, The British Council, and IDP Education Australia: IELTS Australia.

Upshur, J. and T.J. Homburg 1983. Some relations among language tests at successive ability levels. In Oller, J.W. jr (ed.) *Issues in language testing research*. Rowley, MA: Newbury House, 188-202.

Upshur, Jack and Carolyn Turner 1999. Systematic effects in the rating of second language speaking ability: test method and learner discourse. *Language Testing* 16 (1), 82-111.

Van Patten, B. 1990. Attending to content and form in the input: an experiment in consciousness. *Studies in Second Language Acquisition* 12, 287-301.

Wainer, Howard and Henry Brown 1988. Historical and epistemological bases of validity. In Wainer, H. and Brown, H. (eds.) 1988, 1-2.

Wainer, Howard and Henry Brown (eds.) 1988. *Test validity.* Hilsdale, N.J.: Lawrence Erlbaum.

Wainer, Howard and G. Kiely 1987. Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185-202.

Wainer, Howard and Robert Lukhele 1997. *How Reliable Is the TOEFL Test?* TOEFL Technical Reports 12. Princeton, NJ: Educational Testing Service.

Wallace, Craig 1997. IELTS: Global Implications of Curriculum and Materials Design. English Language Teaching Journal, 1997, 51, 4, Oct, 370-373.

Waters, Alan 1996. *A Review of Research into Needs in English for Academic Purposes of Relevance to the North American Higher Education Context.* TOEFL Monograph Series 6. Princeton, NJ: Educational Testing Service.

Way, Walter D., Patricia A. Carey, and Marna L. Golub-Smith 1992. An Exploratory Study of Characteristics Related to IRT Item Parameter Invariance with the Test of English as a Foreign Language. TOEFL Technical Report 6. Princeton, NJ: Educational Testing Service.

Way, Walter D. and Clyde M. Reese. February 1991. *An Investigation of the Use of Simplified IRT Models for Scaling and Equating the TOEFL Test.* TOEFL Technical Reports 2. Princeton, NJ: Educational Testing Service.

Weir, Cyril 1983. Identifying the language problems of overseas students in tertiary education in the United Kingdom. Unpublished doctoral thesis, University of London, London.

Weir, Cyril 1988. *Communicative language testing.* University of Exeter.

Weir, Cyril 1990. Weir, C. 1990. *Communicative language testing*. New York: Prentice Hall.

Weir, Cyril 1993. *Understanding and Developing Language Tests.* New York: Prentice Hall.

Westaway, Gillian, J. Charles Alderson, and Caroline Clapham 1990. Directions in testing for specific purposes. In J.H.A.L. de Jong and D. Stevenson (eds.), *Individualizing the assessment of language abilities.* Clevedon, Avon: Multilingual Matters, 239-256.

Wiley, David E. 1991. Test validity and invalidity reconsidered. In R. E. Snow and D. E. Wiley (eds.) *Improving inquiry in the social sciences: A volume in honor of Lee J. Cronbach.* Hillsdale, NJ: Lawrence Erlbaum.

Widdowson, Henry 1989. Knowledge of language and ability for use. *Applied Linguistics* 10, 128-137.

Wilson, K.E. 1982. *A Comparative Analysis of TOEFL Examinee Characteristics, 1977-1979.* TOEFL Research Reports 11. Princeton, NJ: Educational Testing Service.

Wilson, K.E. 1982. *GMAT and GRE Aptitude Test Performance in Relation to Primary Language and Scores on TOEFL*. TOEFL Research Reports 12. Princeton, NJ: Educational Testing Service.

Wilson, K.E. 1987. *Patterns of Test Taking and Score Change for Examinees Who Repeat the Test of English as a Foreign Language.* TOEFL Research Reports 22. Princeton, NJ: Educational Testing Service.

Wood, Sandra (ed.) 1998. *IELTS International English Language Testing System Research Reports 1998.* EA Journal Occasional Paper 1998 Volume 1. Sydney, NSW: ELICOS Association Limited and IELTS Australia Pty Limited.

Yalow, E. S. and W. J. Popham 1983. Content validity at the crossroads. *Educational Researcher* 12 (8), 10-14.

Yamamoto 1995. *Estimating the Effects of Test Length and Test Time on Parameter Estimation Using the HYBRID Model.* TOEFL Technical Reports 10. Princeton, NJ: Educational Testing Service.

Yin, Robert K. 1994. *Case study research. Design and methods.* Second edition. Thousand Oaks: Sage Publications.

Young, Richard 1995. Conversational Styles in Language Proficiency Interviews, *Language Learning* 45 (1), 3-42.

Appendix 1. TOEFL Research Reports

identified by ETS (1999c:2) as relevant to the Reading section

Report ID	Purpose of study	Materials / data	Methods	Results	Relevant to
RR-1 Clark (1977)	to check that the TOEFL items are not inappropriately difficult or overly sophisticated for non-native speakers by studying native speaker performance	Total and section scores of native speakers (N=88) on two forms of the 3-section TOEFL test, post-test questionnaire on difficulty of sections and item types	descriptive statistics, percentage fails, analysis of items with highest percentage-fail figures	Score ranges consistently high and strongly negatively skewed, clearly distinct from non-native speaker score ranges; some tendency for summarization or inference items in the reading section to be difficult for native speakers	face/content validity, difference variables in examinee performance, decisions/cut scores
RR-2 Pike (1979)	to obtain information useful for evaluating and revising TOEFL content and content specifications, to investigate how many section scores should be reported on the score reports	comparison of five-section TOEFL with scores from four new objective test types, cloze, re-writing, and two productive skills criterion measures: speaking and writing. Student N=442, three national backgrounds	concurrent validation related to six areas of language competence, implemented through factor analysis	Listening was relatively independent and correlated well with speaking; structure correlated well with productive writing and speaking and with section 5 Writing Ability, the two sections could be combined. Vocabulary correlated highly with Reading, recommend combination.	decisions/cut scores, innovative formats
RR-3 Angelis, Swinton, & Cowell (1979)	to compare native and nonnative speaker performance on verbal aptitude tests designed for native speakers, to see analyse NNS differences by TOEFL score ranges	TOEFL scores, GRE verbal scores, SAT verbal scores, Test of Standard Written English scores; graduate and undergraduate students, total N=396	descriptive statistics for each student type, comparison of means with NS performance on aptitude tests, correlations with TOEFL scores	Levels of TOEFL scores at which aptitude test scores begin to be meaningful: 475 for GRE Verbal and 435 for SAT Verbal	concurrent validity, score interpretation, explanation of differences in examinee performance
RR-5 Powers (1980)	to study the relationship between TOEFL and GMAT scores, to find TOEFL threshold scores beyond which GMAT scores begin to be meaningful, to investigate discrepancies across background variables	TOEFL scores and GMAT scores for 5,793 nonnative speakers of English from 26 countries (+ an “other countries” category)	least-squares regression of GMAT total and subscores on TOEFL total and subscores	TOEFL and GMAT are different tests; a minimum score of approximately 450 on TOEFL is required before GMAT verbal scores begin to discriminate among candidates with respect to the kind of verbal ability measured in the GMAT	construct validity, concurrent validity, score interpretation, examinee populations

Report ID	Purpose of study	Materials / data	Methods	Results	Relevant to
RR-6 Powers and Swinton (1980)	to determine the components of abilities that the TOEFL measures, to investigate the nature of minor dimensions	TOEFL total and section scores for groups representing seven language backgrounds, group N 600-1000	Factor analysis (four-factor target matrix), regression of candidate background variables on factor structure for the test	three or four factors appeared necessary for each language group, with both similarities and differences across groups. Listening was separate across groups, whereas reading, writing, vocabulary and grammar were grouped differently across backgrounds.	construct validity
RR-9 Alderman and Holland (1981)	to analyse possible bias of TOEFL items related to examinees' native language background	Item scores on all items of the TOEFL by members six different language groups, group N approximately 1000	chi-square analysis of observed and expected item performance; linguistic explanations for discrepant item performances by experts	Nearly seven eighths of the items were found to be sensitive to the examinees' native languages. Specialists attributed the differences to linguistic similarities between English and the native language. The same reviewers were unable to predict language-specific DIF from inspecting the test and keys without response data.	test/item bias, examinee populations
RR-10 Alderman (1981)	to investigate language proficiency as a moderator variable in testing academic aptitude	total and section scores from the TOEFL, the ESLAT (a Puerto Rican test of English) and from verbal aptitude tests given in the native language (Spanish) and in English. Total N=384	regression of SAT scores on the scores from the aptitude test in Spanish and from the TOEFL and ESLAT	language proficiency is a moderator variable in assessing academic aptitude; appropriacy of aptitude tests in L2 increases if second language proficiency is high. Thresholds suggested: TOEFL 500, ESLAT 600.	construct validity, predictive validity, concurrent validity, score interpretation
RR-11 Wilson (1982)	A Comparative Analysis of TOEFL Examinee Characteristics 1977-1979	TOEFL scores, examinee background information	Cross-tabulation, sub-population comparisons by region, native country, and native language	Nature of TOEFL examinee sub-populations from different backgrounds, TOEFL score ranges for regions	examinee populations (test use)

Report ID	Purpose of study	Materials / data	Methods	Results	Relevant to
RR-12 Wilson (1982)	GMAT and GRE Aptitude Test Performance in Relation to Primary Language and Scores on TOEFL	TOEFL scores, GRE Aptitude Test scores, GMAT test scores. Total N=2,442	comparison of means with NS performance on aptitude tests, correlations with TOEFL scores	aptitude test scores earned by foreign students are mediated by their English language proficiency. Only when TOEFL scores reach approximately 625 do verbal aptitude scores reach the range of NS scores on tests of academic aptitude	construct validity, concurrent validity, score interpretation
RR-14 Swinton (1983)	A Manual for Assessing Language Growth in Instructional Settings	TOEFL scores of students at a one-week and 13-week interval from first test with intervening intensive instruction in English	subtraction of pretest scores from posttest scores, with effects of practice and regression towards mean removed by taking the one-week retest gain into account	Students with initial scores in the 353-400 showed a real gain of 41 points during the 13 weeks of instruction, and students with initial scores at the 453-500 range a 25-point real gain. The lower the initial score, the greater the probable gain on a fixed-length course.	socio-pedagogical impact
RR-17 Duran, Canale, Penfield, Stansfield, and Liskin-Gasparro (1985)	To describe the content characteristics of TOEFL items and sections in terms of communicative competence	Exploratory framework of communicative competence covering competence areas, features of performance required, relevance for academic and social language uses, and minimum mastery level; one TOEFL test form; two researchers completed each analysis	Analysis of the TOEFL test form: domain description based on a communicative skills checklist, preliminary analysis of performance features, evaluation of authenticity, preliminary description of minimum levels of mastery	TOEFL tests a wide range of linguistic and communicative competences, albeit receptively. Long contextual items in reading and writing good communicatively, decontextualized items less so. The test is appropriate for basic, intermediate, and advanced learners.	construct validity, content validity
RR-21 Stansfield (Ed.) (1986)	Toward Communicative Competence Testing: Proceedings of the Second TOEFL Invitational Conference	invited papers and summaries of the discussions	discussion of the theoretical construct of communicative competence and its constituent parts in relation to what is measured in the TOEFL test, and to possible test revisions	various recommendations to change the emphasis on discrete-point testing towards more integrated, more extended-passage, more authentic language use situations. An account of revisions already under way, notably the introduction of the Test of Written English.	construct validity, face/concurrent validity

Report ID	Purpose of study	Materials / data	Methods	Results	Relevant to
RR-22 Wilson (1987)	Patterns of test taking and score change for examinees who repeated the TOEFL test within 24 to 60 months after they first took the test	Background variables and total and section scores on the TOEFL from examinees re-taking the TOEFL at least 24 and not more than 60 months after first administration	descriptive statistics, investigation of differences between scores, proportion of examinees re-taking the TOEFL from different regions	repeaters registered substantial average net gains in performance even at 24 month interval; some national and linguistic groups re-take the TOEFL proportionally more often than others, and	practice/sequence effects (test implementation)
RR-23 Manning (1987)	to investigate the validity and practical utility of cloze-elide tests of English proficiency for students similar to the TOEFL candidate population	scores from TOEFL, cloze-elide tests, cloze tests, and mc-cloze tests gained by students on intensive English programs (N=1,208); essay scores, teacher ratings, student self-ratings, and background information	score intercorrelations, factor analysis, multiple regression; analysis of comparability of sample to standard TOEFL populations	cloze-elide tests compare favorably with more commonly used testing procedures. They exhibit internal consistency estimates of .89 but are administrable in a shorter testing time than a multiple choice test. They load on all TOEFL factors approximately equally, which shows they provide a general measure.	format selection, innovative formats
RR-25 Hale (1988)	Interaction of a student's major-field group with the text content in determining performance on TOEFL reading passages				difference variables in examinee performance, test/item bias, socio-pedagogical impact
RR-26 Hale, Stansfield, Rock, Hicks, Butler, and Oller (1988)	To develop a multiple-choice cloze test; to investigate the relationship between different types of MC cloze items and TOEFL section scores	Scores and sub-scores on multiple-choice cloze test, TOEFL total scores and section scores (N=11,290)	Confirmatory factor analysis, IRT parameter estimation, correlation, multiple regression	No evidence that distinct skills are measured by nonlistening parts of the TOEFL; no evidence that different MC cloze item types assessed different skills (reading, grammar, vocabulary)	format rationale/selection (test construction), innovative formats

Report ID	Purpose of study	Materials / data	Methods	Results	Relevant to
RR-27 Oltman, Stricker, and Barrows (1988)	To assess the influence of examinees' native language and their level of English proficiency on the structure of the TOEFL	Detailed item response data (correct response, incorrect response, omitted, not reached); all TOEFL items, seven language groups and three levels of performance, 400 examinees in each subsample	Multidimensional scaling	TOEFL construct validity is supported, the dimensionality of the TOEFL test and of competence in English depends on examinees' English proficiency. More differentiated constructs are measured for low-scoring examinees	construct validity, diagnostic value, reporting/scaling;
RR-28 Boldt (1988)	Whether there are examinee subgroups whose TOEFL performance is explained by different latent structures than the rest	TOEFL total and section scores, examinee background variables (money available, major subject, native language, national origin, gender) (N=94,000)	Factor analysis, regression analysis, analysis of variance	a single factor (group) gave a very accurate accounting for the proportions of joint item success, ie. latent group effects are small	construct validity, sample dimensionality (examinee performance)
RR-29 Angoff (1989)	Whether TOEFL examinees tested in their native countries are disadvantaged because of American references in the test		Mantel-Haenszel analysis	TOEFL does not place foreign-tested examinees at a disadvantage	test/item bias
RR-30 Secolsky (1989)	to determine whether the TOEFL test is speeded according to established criteria		(two exploratory approaches)	Section 3 pretest administrations may be slightly speeded; further confirmation is needed	satisfying assumptions (test use), testing time (test implementation)
RR-31 Hicks (1989)	development of an experimental TOEFL computerized placement test using conventional scoring methods		a testing algorithm that routed examinees through item blocks or testlets and permitted backtracking to review answers and change them		computer-adaptive testing
RR-32 Hale, Rock, and Jirele (1989)	To examine the factor structure of the TOEFL test	TOEFL total and section scores; domestic and overseas populations, five language groups, low- and high-scoring groups (N=20,000)	Confirmatory Factor Analysis	two factors, one associated with the Listening Comprehension section, and the other with the rest of the sections. Factor structure consistent across all groups	construct validity

Report ID	Purpose of study	Materials / data	Methods	Results	Relevant to
RR-35 Henning (1991)	comparative functioning of eight multiple-choice vocabulary item formats; differences: length and inference-generating quality of stem, nature of task, and degree of passage embeddedness of item stems or response options	Responses to 1,040 vocabulary test items, by 190 adult ESL students. Self-reports of prior familiarity with item types.	Estimates of item difficulty, item discriminability, criterion-related validity, and subtest reliability	Use of vocabulary embedded in reading passages, as well as use of vocabulary stems with inference-generating information, resulted in superior item functioning. Some item types were more familiar than others, and the most familiar item types showed a positive correlation with successful performance on the item type.	construct validity, response validity, format rationale/selection (test construction), innovative formats
RR-36 Henning and Cascallar (1992)	Nature of Communicative Competence, interrelations among its variables, relationship to existing TOEFL test; to propose a tentative construct model for TOEFL	major variables of communicative competence from theoretical literature, ratings of 79 students on experimental tasks; same students' TOEFL, TSE, and TWE® scores	Analysis of theoretical literature, analysis of academic communication situations, creation of experimental tasks, scoring of experimental performances; descriptive statistics, ANOVA, multiple correlation and regression	General: five-minute samples required for oral communicative ratings, approx. 15 minutes for similar writing samples; oral performance improves in later tasks; communicative proficiency is situation specific (context, purpose, function). TOEFL-specific: detailed framework needed for test construction; traditional measures of structure are not empirically unrelated to communicative performance measures; fluency of cognition and strategic competence worthy of measurement	construct validity, score interpretation, underlying processes
RR-41 Boldt, Larsen-Freeman, Reed, and Courtney (1992)	to align ACTFL Proficiency descriptions of test takers' language performance with TOEFL section score ranges	ESL instructors' ratings of their students' listening, reading, and writing proficiency on the ACTFL scale, the same students' TOEFL section scores; N per skill 400 – 600	quantification of ACTFL descriptors, rating, investigation of adjustments for severity, reliability of ratings, cross-tabulation, correlation of ratings	Correlations, limited by reliabilities, were substantial although could have been bigger. Thus ACTFL ratings and the TOEFL test tap similar underlying skills, and ACTFL descriptive scales can be used to an extent in interpreting TOEFL scores. Percentile distribution tables provided.	predictive validity, score interpretation, performance descriptors

Report ID	Purpose of study	Materials / data	Methods	Results	Relevant to
RR-44 Freedle and Kostin (1993)	Prediction of TOEFL Reading Comprehension Item Difficulty for Three Item Types: Main Idea, Inference, and Supporting Idea	213 TOEFL reading comprehension items, 98 passages, scored responses from 2000 examinees; textual variables related to reading passages, items, and text-item overlap	ANOVA, MANOVA, stepwise regression	33 to 61 percent of item difficulty variance can be accounted for by a relatively small number of variables related to features of texts and text-item overlap, ie difficulty is related to *text* difficulty. Nesting may be a problem.	construct validity
RR-45 Henning (1993)	comparative global and component estimates of reliability; test-retest change in subtest difficulty within short time (eight days)	Component and total TOEFL scores for examinees for two test administrations with a time interval of eight days	test-retest, alternate form, and internal-consistency reliability	test-length-adjusted reliability estimates were found to be adequately high across reported components and total test scores; the study was limited by a small sample size that were not perfectly representative of the TOEFL examinee population in language background and mean proficiency	language acquisition/loss (examinee performance), internal consistency, alternate forms and test-retest reliability
RR-47 Schedl, Thomas, and Way (1995)	Assessment of speededness of TOEFL Section 3 if vocabulary items are embedded in reading passages	Examinee (N=1300) score patterns to an institutional TOEFL and three experimental TOEFL Section 3 tests, different maximum test times and numbers of items; comparative analyses based on 47 common items	Assessment of sampling procedures, traditional assessment of speededness, outlier analysis, score comparisons, equating analyses, alternate form reliability	Implementation of revised Section 3 consisting of five reading passages with a total of 50 items was supported, no less than 55 minutes should be allowed. Additional passages induce greater speededness effects than additional items. Current TOEFL scale can be maintained with revised test.	construct validity, satisfying assumptions (test use) selection of test format, decision of component length, testing time (test implementation)
RR-53 Schedl, Gordon, Carey, and Tang (1996)	to investigate the dimensionality of the TOEFL reading test: do “reasoning” items measure something different from the rest of the TOEFL items?	Reading scores from ten different TOEFL administrations, each with more than 1000 examinees, categorised by skill tested (reasoning/other)	Stout's procedure for assessing essential unidimensionality, and NOHARM nonlinear factor analysis, to investigate a hypothesized two-factor model	TOEFL reasoning items cannot be shown to measure a unique construct. Two factors *were* found, however, and exploratory analyses indicated that passage content or position may be the cause	construct validity

Report ID	Purpose of study	Materials / data	Methods	Results	Relevant to
RR-57 Boldt and Courtney (1997)	to investigat the ways in which TOEFL test scores are used by colleges and universities, to examine the ways in which they set standards	Institutions' responses to a questionnaire on the use of TOEFL scores when selecting international students	Survey which focused on minimum standards used by institutions, the ways in which the standards were established, and steps in decision making process when admitting international students	Institutional standards most commonly based on the practices of other institutions rather than local research. Use of cut scores as rigid standards of admission was rare; instead additional measurement or English training was required for those falling below an institutional minimum.	decisions/cut scores, examinee/user populations

Appendix 2. TOEFL Technical Reports

identified by ETS (1999c:2) as relevant to the Reading section

Report ID	Purpose of study	Materials / data	Methods	Findings	Relevant to
TR-1 Oltman and Stricker (1991)	to investigate the possibility of developing alternative score reporting scales for the TOEFL test, two types of scoring	Detailed TOEFL response data (correct, incorrect, omitted, not reached), and right/wrong scoring data for all sections	multidimensional scaling	Both scoring types produced clusters of items in the test sections; these clusters might be more homogeneous and more distinct than their parent sections, and thus better suited for diagnostic use. Cluster patterns for the two scoring types were only different for extreme scoring students.	construct validity, reporting/scaling (test information)
TR-2 Way and Reese (1991)	to explore the use of one-parameter and two-parameter IRT estimation models for scaling and equating the TOEFL test instead of the three-parameter logistic model currently used	Artificial data used for simulating typical TOEFL equatings. Four simulated sample sizes: 600, 900, 1200, and 1500 responses per equating set, giving total sample sizes of 2400, 3600, 4800 and 6000.	One-parameter, two-parameter and three-parameter logistic models	Use of three-parameter model was supported. Discrepancies between score conversions tended to occur at the lower and upper ends of the score scales. Quality of simulated equatings based on the three-parameter model did not appear to be sensitive to different sample sizes.	reporting/scaling (test information), equating (test construction)
TR-4 Boldt (1991)	whether the proportional item response curve (PIRC) model could serve as a basis for simpler equating methods than are currently used by the TOEFL program	Item response curves based on pretest data, actual item response curves and actual score means and standard deviations. Rules for prediction and for comparison of methods.	Prediction of item responses, test scores, and test score means and standard deviations using PIRC, a three-parameter logistic model, and a modified Rasch model. Comparison of models.	Predictions made by all the models were approximately equally accurate. Size of estimation sample seemed to make little difference.	equating (test construction)

Report ID	Purpose of study	Materials / data	Methods	Findings	Relevant to
TR-5 McKinley and Way (1992)	whether listening, structures, written expression, vocabulary and reading could be distinguished in score data using IRT	examinee responses to two forms of the TOEFL test, 146 items in each, 5000 examinees, stratified random sample	unidimensional item response theory (IRT), exploratory multidimensional IRT (MIRT), and confirmatory multidimensional IRT (CMIRT) models applied on data	The same ability structure that had been found in previous factor analytic studies was found with MIRT and CMIRT. The test is strongly unidimensional. On secondary dimensions, listening is most clearly distinguishable. Some tendency in foreign samples to show a distinct reading dimension as well. The consistent Akaike index is a useful criterion for comparing different models.	construct validity, sample dimensionality (examinee performance)
TR-6 Way, Carey, and Golub-Smith (1992)	to explore item features that may contribute to a lack of IRT item parameter invariance	data and IRT item parameter estimates from seven TOEFL final forms	fit of operational data with pretest item parameter estimates, and reestimated item parameter estimates with pretest item parameter estimates	item position changes and prolonged intervals between pretest and final form administration may contribute to lack of IRT item parameter invariance	satisfying assumptions (test use), item pretesting/selection (test construction), test-retest reliability
TR-7 Tang, Way, and Carey (1993)	to compare the performance of LOGIST and BILOG on TOEFL IRT-based scaling and equating	both real and simulated data, two calibration structures, different sample sizes	fit of operational data with pretest item parameter estimates, analysis of root mean squared error statistics	item parameter estimates obtained from the smaller real data sample sizes were more consistent with the larger sample estimates when based on BILOG than when based on LOGIST. Pretest sample sizes be at least 1,000 for LOGIST should be retained if at all possible	equating (test construction)
TR-8 Boldt (1994)	to "equate the test to itself" using the product of a person parameter and an item parameter rather than the logistic curve	two samples of responses to identical item sets, various sample sizes	comparison of equating results for the sections of the TOEFL test using variations of sample size and anchor test difficulty, assessment with mismatched samples through selection on a correlated variable	The largest discrepancies between scores identified as comparable occurred for the logistic-based models at the lower extreme scores, and for the simple models at the upper extreme score.	equating (test construction)

Report ID	Purpose of study	Materials / data	Methods	Findings	Relevant to
TR-9 Chyn, Tang, and Way (1995)	to investigate the feasibility of the Automated Item Selection (AIS) procedure for the Test of English as a Foreign Language	Item pools of varying sizes from 290 to 1432 items. Statistical test construction rules based on IRT indicators. Test development based construction rules based on content considerations.	Two TOEFL final forms were assembled using AIS with statistical and content criteria. Tests were evaluated on statistical and content-related criteria.	Statistical consistency (parallelism) of the tests assembled using AIS appeared to be superior to the consistency of tests assembled using traditional test assembly procedures. Visible gains in time efficiency in item selection for Sections 1 and 2 and the potential for time gains in Section 3.	item pretesting/selection (test construction), machine test construction, item banking
TR-10 Yamamoto (1995)	to investigate alternative indicators of test speededness	detailed response data (correct answer, incorrect answer, omitted, not reached)	Extension of HYBRID model to determine when each examinee switches from an ability-based response strategy to a strategy of responding randomly. Test speededness was evaluated by estimating proportions of examinees switching at all possible points in the test.	Estimated IRT parameters based on the HYBRID model were found to be more accurate than those based on ordinary IRT analysis. The proportion of examinees who were affected by speededness of the test at 80 percent test completion was nearly 20 percent. For this group, responses on the last 20 percent of items did not represent the examinees' true ability.	satisfying assumptions (test use), testing time (test implementation)
TR-11 Boldt and Freedle (1996)	to improve the predictions of item difficulty by using a nonlinear process (neural net in prediction and genetic algorithm in choice of variables)	data from Freedle and Kostin (1993) (TOEFL RR-44): 213 items nested in 100 reading paragraphs and a reduced 98 non-nested item sample, 75 content descriptors, equated item deltas	comparison of linear prediction and neural net prediction, use of linearly selected item characteristics and genetic algorithm-selected item characteristics	Neural net only improved prediction when prediction variables were selected by genetic algorithm. Only two variables, both related to text-item overlap, were the same as Freedle and Kostin's. Technical reasons may explain differences between nested and non-nested and low- and high-scoring examinees. Neural net is an experimental method for predicting difficulty and may suffer from capitalization on chance.	construct validity, score interpretation, item pretesting/selection (test construction)

Report ID	Purpose of study	Materials / data	Methods	Findings	Relevant to
TR-12 Wainer and Lukhele (1997)	How Reliable Is the TOEFL Test?	scores on four forms of the TOEFL test	reliability estimation using a hybrid IRT model	Very little difference in overall reliability when testlet items were assumed to be independent and when their dependence was modeled. A larger difference when various test sections were analyzed individually. Up to 40 percent overestimate in reading testlets, with longer testlets showing the most local dependence. The test was unidimensional enough for the use of univariate IRT to be efficacious.	Internal consistency reliability

Appendix 3: Published reports and studies related to the initial development and validation of IELTS

Author, date	Focus	Materials and methods	Main findings	TD / VAL
Alderson 1988	Progress report on IELTS development, description of stages of development, plans for writing specifications, connection to theories and practices of test development and validation	Participant knowledge of IELTS development, analysis and evaluation of ESP testing and limits of needs analysis, reporting of development rationale, analysis of purposes of test specifications	Limitations of needs analysis led IELTS team to use stakeholder feedback and iterative commentation and revision as main guidelines in test development. Specifications and tasks developed simultaneously. Specifications serve two aims: construct definition and guidance of test and task writing.	TD: stages of test development, specification writing in iteration with task writing VAL: iterative specification and task writing as content validation
Westaway, Alderson and Clapham 1990	First stage of IELTS development: reactions to ELTS from test administrators, receiving institutions, British Council staff, and language testers and teachers	Questionnaires to and interviews with stakeholders, commissioning of papers by language testers, test taker report forms; descriptive statistics and conceptual summary	Number of specific modules could be reduced, admissions officers use overall scores but departments like score profiles, administrators find the test cumbersome, testers agreed Munby was outdated but could not propose replacement	TD: theoretical and empirical background research, use of stakeholder opinions and expectations
Alderson 1991	Discussion of nature of assessment scales, description of IELTS scale development	Examples of scales used by others, existing ELTS scale; developer account of rationales	Three functions of scales: user-oriented, assessor-oriented, and constructor-oriented. Iterative scale development: stakeholder comments and trial assessments used as input.	TD: nature of scales, scale development
Alderson and Clapham 1992	Definition of the construct of language ability in IELTS, implications for operationalization in test specifications	Survey of language testers' views on appropriate constructs to inform IELTS development, analysis of responses, drawing of implications	No agreement among applied linguists to replace Munby's model, though views of language as communicative and contextualised were stressed. An eclectic model had to be used to stay on project schedule.	TD: how to define constructs for tests VAL: test can at best be an indirect operationalization of theory
Alderson and Clapham (eds.) 1992	Report/overview of first stage of develoment (as above); proposals for the design of IELTS and report on decisions	Summary of results from questionnaires, interviews and commissioned papers, summary of conference discussions, report on development decisions	No screening test to be developed. Listening and speaking to be common for all, reading and writing to include three academic specialisations. Need for general training module identified.	TD: operational decisions

Author, date	Focus	Materials and methods	Main findings	TD / VAL
Alderson 1993	Relationship between grammar and reading, rationale for development decision to drop grammar test	pretest data from samples of 195-842 learners depending on module, reliability coefficients, correlations of section scores, factor analysis	Reading and grammar were closely related on all indicators. Need to report skill scores and need to shorten test suggested deletion of grammar.	TD: decisions on test format based on empirical data VAL: what was the test testing?
Ingram and Wylie 1993	Structure of IELTS speaking test, rationale for its structure, nature of skills assessed, discusison of proficiency assessment scales	Participant knowledge of development of IELTS speaking section, theoretical knowledge of assessment scales, argument	Documentation of rationale for IELTS speaking test, research agenda for future development and validation of the speaking module	TD: test description, research agenda VAL: research agenda
Clapham 1993	Bacground knowledge in testing reading: do students score significantly higher on a reading test in their own area?	Groups of 155-174 students took two IELTS reading modules; MANOVA of mean scores to see effects of background knowledge	Students not significantly disadvantaged even if they take a test outside their academic discipline. Background knowledge is more complex than future area of study	TD: should there be specific purpose modules in the test VAL: is ESP testing justified
Clapham 1996a	Effect of background knowledge on reading comprehension	examinee scores (N=204-328) on two IELTS reading sections, data on their background knowledge and intended discipline, specialist analysis of specificity of text content; MANOVA, specificity rating through task characteristics	In general, students do better in their own area. However, students may be significantly disadvantaged if a text is highly subject-specific even if it is in their own area. There may be a proficiency threshold below which background knowledge does not benefit students.	TD: test specificity: both specific and generic tests possible VAL: validity of scores as indicators of subject specific ability?
Clapham 1996b	Content analysis of tasks in three IELTS modules to detect reasons for subject specificity of texts	(as above)	(as above)	TD&VAL: validity of developing subject-specific reading tests
Clapham and Alderson (eds.) 1997	six reports of the development and trialling stages of different IELTS modules, rationale of IELTS scale and scoring, analysis of pretest data	Participant knowledge of IELTS development, records and rationales of development considerations and decisions, reliability & IRT indicators for pretest items	an iterative chain of trialling and revision of items and specifications led to satisfactory development solutions, relationship between test-based evidence and descriptors on reporting scales is somewhat uneasy but useful, internal consistency and validity coefficients of pretests adequate	TD: trialling, revision, pretesting, data analysis, comparability across cultures VAL: measurement quality, score comparability

Appendix 4: Published reports and studies related to the operational development and validation of IELTS

Author, date		Focus	Materials and methods	Main findings	TD / VAL
Charge and Taylor 1997		Progress report on IELTS development, description of revision in 1995, rationale for revision	Knowledge of IELTS development from testing board's point of view, reporting of rationale	One generic academic test was introduced because of Clapham's validation research and difficulties in assigning students to modules. Reading and writing disconnected to avoid undesirable variation in performance strategies	TD: development rationale
Wood (ed.) 1998		IELTS Australia-sponsored research on IELTS: Volume 1 (6 papers)			
1	Brown and Hill 1998	Interviewer style and its effect on candidate performance in IELTS oral interview	32 candidates interviewed by 2 of 6 interviewers; discourse analysis of strategies of two 'easy" and two 'difficult' interviewers	Interviewer style varies. The board must decide which style it wants, train, encourage self-monitoring, and monitor interviewers.	TD: improving the quality of testing procedures VAL: comparability of skills tested
2	Brown 1998	effect of instruction on IELTS writing scores	comparison of IELTS and non-IELTS writing course on IELTS writing scores, student N 9 on IELTS and 5 on general.	Participants on IELTS course improved their score more than participants on general academic writing course. Whether IELTS course made students better academic writers is questionable.	VAL: impact
3	Coleman and Heap 1998	possible misinterpretation of rubrics in listening and reading modules	examinee responses to reading and listening sections (N=40-115), 11 post-test interviews	Rubrics were not unclear, examinees understood what to do. Wording of some test questions could be improved, marker reliability should be monitored	TD: improving the quality of testing procedures
4	Cotton and Conrow 1998	predictive validity of IELTS at University of Tasmania	correlation between IELTS total and skill scores with academic GPA, staff ratings and student self-ratings, self-reported difficulty and use of language support. Student N=33	Correlations with academic performance not significant, reading and writing correlated weakly (.34 - .46) with staff ratings and student self-ratings. Small sample size limits usefulness of study.	VAL: score use

Author, date		Focus	Materials and methods	Main findings	TD / VAL
5	McDowell and Merrylees 1998	the degree to which Australian institutions of higher education use IELTS as their language criterion, their opinion on IELTS usefulness	survey of all institutions, personal interview with "a number of academic staff at a range of universities across Australia"	IELTS is the most commonly used test as an admission criterion, and users are satisfied with its quality. Profile of IELTS was raised.	VAL: acceptability
6	Mok, Parr, Lee and Wylie 1998	comparability of IELTS scale with ACCESS scale, with ASLPR as a linking device to establish comparison	total and section ratings for 355-759 examinees rated with one of the scales, those rated with more than one scale numbered 32 altogether	Subskill-specific scales within each examination system were different, which meant that comparability should not be established between overall scores but between section scores	VAL: score comparability
Tulloh (ed.) 1999		IELTS Australia-sponsored research on IELTS: Volume 2 (4 papers)			
1	Merrylees and McDowell 1999	attitudes of oral examiners to IELTS oral interview, analysis of transcripts to study variation in test administration	survey of examiners, counts of examiner/examinee turns, words and test length in minutes for each test section	Majority of examiners comfortable with interview. Considerable variation between examiners in amount of speech and lenth of time spent per test section.	TD: improving the quality of testing procedures
2	Celestine and Cheah 1999	effect of student background in science or arts stream in Malay secondary education on performance on IELTS	IELTS total and section scores, examinee background information, comparison of mean scores within proficiency bands	No statistically significant differences overall, but intermediate and weak students did better if they came from the science track in Malay schools.	VAL: score comparability TD&VAL (potential): score explanation
3	Hill, Storch and Lynch 1999	effectiveness of IELTS and TOEFL as predictors of academic success at University of Melbourne	IELTS (N=35) or TOEFL (N=27) scores and first semester course grades, questionnaires (N=66) and interviews (N=22)	IELTS prediction was moderate (.540) while TOEFL prediction was weak (.287), TWE scores not included in TOEFL values. Those with lowest grade point averages sought language support.	VAL: score use. Examinations help identify students who need language support.
4	Moore and Morton 1999	authenticity of IELTS writing task 2 against university writing assignments	analysis of 20 task 2 items and 155 university writing assignments with discourse analysis categories, survey of 20 staff members regarding comparison between their writing assignment and task 2 items	Task 2 items correspond to university writing assignments in terms of genre. Use of own ideas instead of sources and research techniques common on test, topics often concrete and rhetorical functions restricted. Propose re-introduction of link between reading and writing.	TD&VAL: authenticity of tasks, possibility to increase authenticity through modification of specifications and small re-design of test

Appendix 5: Published reports on the conceptual development of TOEFL 2000

#, author, date	Focus	Materials and methods	Findings/contributions	TD / VAL
MS-1 Ginther and Grant 1996	academic needs of native English-speaking (L1) college students in the United States from several perspectives: student abilities, writing across curriculum, student perceptions	Review of literature on native speaker language requirements, discussion of implications from the perspective of TOEFL 2000	Questions about relevance of survey of native speaker needs, questions about the identification of the appropriate testing domain, the appropriate level of specification of test tasks, the fairness of testing academic tasks, and authentic language use in testing.	a preliminary step for a study examining the academic language needs of entering undergraduate and graduate students in the United States
MS-2 Tang 1996	possible calibration methods for TOEFL 2000 if extended responses are used in combination with selected response items	Review of literature on polytomous scoring models concentrating on empirical trials and comparisons, introduction of PARSCALE analysis program	two commonly used polytomous IRT models are possible for TOEFL 2000: (1) the generalized partial credit model and (2) the graded response model. PARSCALE allows the concurrent calibration of dichotomously and polytomously scored items.	Possibilities of using currently known calibration methods if dichotomously and polytomously scored items are combined in TOEFL 2000
TR-13 Tang and Eignor 1997	possible calibration methods for TOEFL 2000 if extended responses are used in combination with selected response items	experiment with data from existing TOEFL, TWE and TSE tests. Three test forms with reading+TWE (N=1500), two test forms with listening+TSE (N=434, 502). 2PL and 3PL IRT models, gneralized partial-credit and graded response models, analysis using PARSCALE	All five analyses yielded a dominant first factor, indicating that calibration of the two skills combinations was possible. Both the generalized partial credit model and the graded response model could be used, although more detailed analyses were possible with PARSCALE when generalized partial credit model was employed.	Verification of possibility to use currently known calibration methods if dichotomously and polytomously scored items are combined in TOEFL 2000
MS-3 Carey 1996	psychometric and consequential issues involved in the use of performance assessments in high stakes test	Review of literature on psychometric implications of using performance assessment in high stakes contexts, implications for TOEFL 2000	(1) task-specific variance is an issue, reliability lower than in traditional assessments. (2) variance due to raters or interactions of raters with examinees can be reduced with careful training. (3) long and complex performance based tasks are particularly context-bound and their scores of limited generalizability.	Measurement implications of long, complex performance based tasks

#, author, date	Focus	Materials and methods	Findings/contributions	TD / VAL
MS-4 Hudson 1996	Academic reading for TOEFL 2000, with special attention on context and processing	Review of literature on academic reading and on reading processes in L1 and L2. Implications for assessing academic reading in TOEFL 2000.	Recommendations for TOEFL 2000. Expansion beyond selected-response formats recommended but selected response need not be rejected altogether. Tasks should be contextualised and more authentic. Thematically organised parts could be considered. Descriptive score reporting scale should be developed. Reading could be combined with other skills in literacy tasks.	Review of existing literature to draw test development proposals for the Reading section.
MS-5 Hamp-Lyons and Kroll 1996	Academic writing from a communicative competence perspective and composition models perspective, framework development for TOEFL 2000	Different approaches in existing literature to assessing writing. Construction of beginnings of a writing framework for TOEFL 2000 including directions for prompt development, scoring procedures, score reporting, and score use. Consideration of costs, practicality, and possible washback from the test.	Recommendations for TOEFL 2000. The test should have more than one task and task type. Test taker choice recommended. Graduates and undergraduates should be offered different tests. Rater training should be further researched, especially ESL / non-ESL raters. Merits of more fully articulated scoring procedures should be analysed. Rater harshness should be modelled. Multiple forms of score reporting recommended.	Use of existing literature to begin to develop an assessment framework for Writing
MS-6 Waters 1996	Review of research into needs in English for academic purposes relevant to TOEFL 2000	Detailed review of American and British research into EAP needs of students, organised into four-skills needs and other, and analysing source of information, number of informants, level (ug/pg), subject areas, and research methods	Existing research on EAP needs does not form an adequate basis for test construction for American academic contexts. A research program is proposed: study actual language use tasks, analyse the linguistic demands of the tasks, triangulate data, and take into account both language “needs” and student “wants”.	Analysis of existing literature to assess its suitability as a basis for test construction; proposal for a research program

#, author, date	Focus	Materials and methods	Findings/contributions	TD / VAL
MS-8 Douglas 1997	Nature of the construct of speaking ability, influence of testing methods on assessment of speaking, background for decisions on test revision	Review of psycholinguistic processing research, presentation of a speech production model. Review of research on test method as situational and discourse context and their influence on rating, leading to development proposals	A large number of processing and test method characteristics that the developers of the speaking test could take into account. Proposals for TOEFL 2000: concentrate on middle level of contextualisation, integrate speaking and listening. Study of strategic competence particularly important.	Research background for development of a test of speaking that takes individual processing and contextual features into account
MS-10 Chapelle, Grabe and Berns 1997	The TOEFL Committee of Examiners' (COE) model of language ability to guide the development of theoretical construct definitions for TOEFL 2000	References to literature on communicative competence, presentation and report on foundations of COE Model, development of guidelines for construct definition work, discussion of validation framework for TOEFL 2000	Test development should begin from specification of academic contexts of language use to hypothesize what the abilities of interest may be for any specific context. Validation will follow Messick's definition and cover score interpretation and consequences of use.	Record of discussions, foundation for construct definition and for future validation work
MS-15 Bailey 1999	summary of research on language testing washback, proposals for validation plans to study test impact	theoretical literature on language testing washback, development of proposals for validation studies for TOEFL 2000 that concentrate on washback and possible negative impact	Washback can influence participants, processes and products in learning environments. Student perceptions should be studied in addition to teachers. Observations, interviews, questionnaires and discussions should be used in triangulated designs.	Review of literature to develop validation plans for TOEFL 2000, especially concerning washback
MS-16 Jamieson, Jones, Kirsch, Mosenthal, and Taylor 2000	a preliminary working framework for the development of the TOEFL 2000 test to guide the development of more specific skills-based frameworks and research agendas	Record of developments and presentation and discussion of a working framework for the whole project, discussion of rationale for the type of construct definition advocated, summary of related work in literacy assessment, development of example with current TOEFL Reading section	To define constructs for TOEFL 2000, the working groups for reading, writing, listening, and speaking must identify variables that characterise tasks. These must define the situational context, the textual properties of the task and the rubric, i.e. the operations required of examinees to perform successfully. Further, the groups must identify research agenda to validate the variables and use the variables to build an interpretive scheme for scores.	Record of historical background, presentation of working framework, definition of tasks for working groups to define important variables, validate the variables, develop prototype tasks and scoring mechanisms

#, author, date	Focus	Materials and methods	Findings/contributions	TD / VAL
MS-17 Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt, and Schedl 2000	TOEFL 2000 Reading section: definition of working framework	Use of general TOEFL 2000 framework and review of research to build a working framework for the Reading section.	Definition of variables, presentation of sample texts and initial proposals for task types, consideration of technical feasibility and desirables, presentation of research agenda, considerataion of contributions to existing test.	Definition of variables for reading section, progress towards development of specifications and prototype tasks
MS-20 Butler, Eignor, Jones, McNamara and Suomi 2000	TOEFL 2000 Speaking section: definition of working framework	Use of general TOEFL 2000 framework and review of research to build a working framework for the Speaking section	Initial characterisation of variables for both tasks and scoring for a likely semi-direct test, finding that research background did not support test development, extensive research agenda, consideration of technical feasibility and potential contributions	Initial characterisation of variables for speaking section, scoring considerations, call for more research