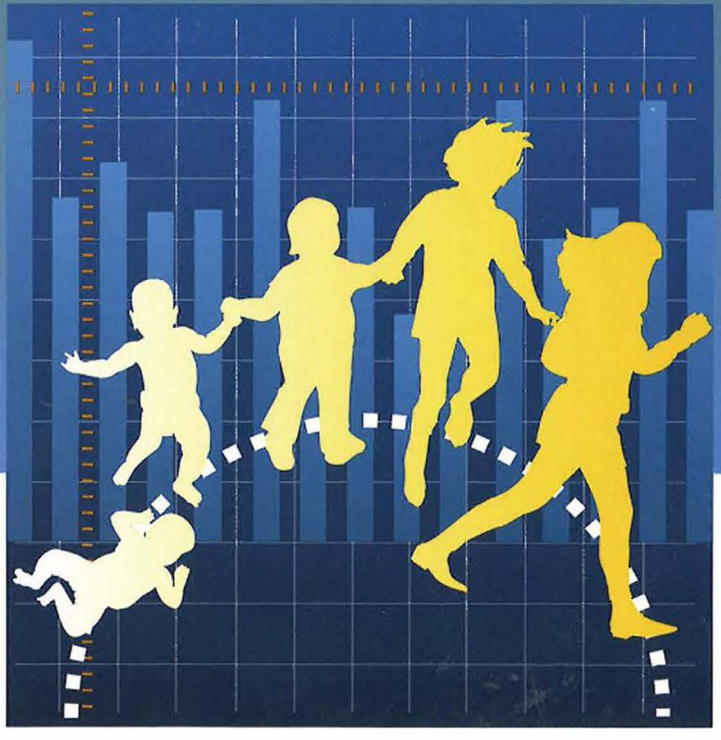
Jaana Ahonniska-Assa

Analyzing Change in Repeated Neuropsychological Assessment



JYVÄSKYLÄ STUDIES IN EDUCATION, PSYCHOLOGY AND SOCIAL RESEARCH 173

Jaana Ahonniska-Assa

Analyzing Change in Repeated Neuropsychological Assessment

Esitetään Jyväskylän yliopiston yhteiskuntatieteellisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston vanhassa juhlasalissa (S212)
joulukuun 9. päivänä 2000 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Social Sciences of the University of Jyväskylä,
in Auditorium S212, on December 9, 2000 at 12 o'clock noon.

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ 2000

Analyzing Change in Repeated Neuropsychological Assessment

JYVÄSKYLÄ STUDIES IN EDUCATION, PSYCHOLOGY AND SOCIAL RESEARCH 173

Jaana Ahonniska-Assa

Analyzing Change in Repeated Neuropsychological Assessment

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ 2000

Editors
Tapani Korhonen
Department of Psychology, University of Jyväskylä
Pekka Olsbo and Marja-Leena Tynkkynen
Publishing Unit, University Library of Jyväskylä

Cover: Pirjo Leirimaa

URN:ISBN:978-951-39-7943-0
ISBN 978-951-39-7943-0 (PDF)
ISBN 0075-4625

ISBN 951-39-0841-0
ISSN 0075-4625

Jyväskylä University Printing House,
Jyväskylä and ER-Paino Ky, Lievestuore 2000

ABSTRACT

Ahonniska-Assa, Jaana
Analyzing change in repeated neuropsychological assessment
Jyväskylä: University of Jyväskylä, 2000, 68 p.
(Jyväskylä Studies in Education, Psychology and Social Research,
ISSN 0075-4625; 173)
ISBN 951-39-0841-0
Yhteenveto: Muutoksen arviointi neuropsykologisessa toistomittauksessa
Diss.

The present study focused on analyzing effects of repeated assessment on performance of several neuropsychological tests. The participants of the study were two groups of children (11.6 and 7.7 years old in the beginning of the research). The neuropsychological tests were repeated nine times with the test-retest interval of two months. The study analyzed in detail the reliability and stability of various scores of the Tower of Hanoi Test in repeated assessment, and suggested revisions in the scoring system of the test. The study also provided growth curves for the Developmental Test of Visuo-Motor Integration, the Underlining Test, the Porteus Mazes Test and the Tower of Hanoi Test. Special interest was focused on the effect of age, cognitive level, and alternative versions on the magnitude of practice effects. Additionally, the study provided an example of an intervention research evaluating change in performance with pre-post assessments, behavior questionnaires, and time-series analysis. The results of the time-series analysis were compared with the growth curve of the control group. The results showed that the visuo-motor task was not sensitive to practice effects, but the visual discrimination and executive function tests were. The alternative versions did not eliminate practice effects. The older participants showed larger practice effects than the younger ones. These findings suggest that reliable conclusions in time-series analysis require minimizing practice effects or separating practice effects from development and intervention effects. If minimizing practice effects is not possible, growth curves help assessing the magnitude of practice effects and separating it from intervention effects.

Keywords: repeated assessment, practice effects in neuropsychological assessment, analyzing change, intervention, Tower of Hanoi

Author's address

Jaana Ahonniska-Assa
Niilo Mäki Instituutti
PL 35
40351 JYVÄSKYLÄ

Supervisors

Professor Heikki Lyytinen
Department of Psychology
University of Jyväskylä, Jyväskylä, Finland

Professor Timo Ahonen
Niilo Mäki Institute
University of Jyväskylä, Jyväskylä, Finland

Tuija Aro, Ph.D
Niilo Mäki Institute
University of Jyväskylä, Jyväskylä, Finland

Reviewers

Docent Juhani Lehto
University of Helsinki, Helsinki, Finland

Docent Eeva-Liisa Helkala
University of Kuopio, Kuopio, Finland

Opponent

Professor Robert McCaffrey
University at Alabama
University of New York State, USA

ACKNOWLEDGEMENTS

First of all I want to express my gratitude to Professors Heikki Lyytinen and Timo Ahonen for awakening my enthusiasm in neuropsychology and neuropsychological research, and for providing supervision and warm support which have helped me to finish this research.

I am very grateful to my third supervisor, Dr. Tuija Aro, for being a colleague and friend for many years. During the last two years of this research, she generously devoted a great deal of time and thinking in reading and commenting on the manuscripts in detail. She has also helped me to overcome the numerous problems involved in writing a thesis in another part of the world where there were few possibilities for face-to-face guidance.

I feel deep gratitude to Asko Tolvanen, M.Sc., for his statistical guidance, for his willingness to devote time and effort to solving my statistical problems, and for his everlasting patience and friendliness in enlightning me on the mysteries of statistics.

I also am grateful to Hanna Mäntynen, Psych.Lic., for collecting and scoring a great part of the data and sharing the difficulties of data collection.

Ms. Terttu Surakka and Maria Haakana, M.A., rescued me from numerous practical problems in filling out grant applications, in meeting timetables, and in getting references and manuscripts mailed to correct addresses in time. Thank you.

I also want to thank my colleagues in Niilo Mäki Institute who have for many years discussed various research problems with me, read my manuscripts and provided me with new ideas.

I want to express my deep gratitude for the friendly and flexible co-operation of the staff of the Special School of Huhtarinne. Their openness to research projects and ability to create a warm and welcoming atmosphere in the school deserves special thanks. Similarly, I am very grateful to the teachers of the Primary School of Muurame, for their supportive attitude and willingness to participate in this research and fpr their friendly patience during the time consuming research project.

This research would not have been possible without the positive attitude and co-operation of the pupils, and their parents and teachers who participated in this research for more than a year and half. I owe them my deepest gratitude.

I feel deep gratitude to my parents, Kerttu and Veikko Ahonniska, who have always believed in me and given me practical support whenever needed. I also want to thank my mother-in-law Rachel Assa and my sister-in-law Nurit Gottlieb for their day-to-day assistance while I have been trying to balance motherhood with research..

I owe a great debt to my husband, Yoav Assa, who saved me from countless hours of work and endless frustration by providing badly needed help in uncountable technical and graphical problems. He also has warmly and patiently supported me during all these years.

I also want to thank Ella and Netta for helping me to find balance between research work and family, and for increasing my satisfaction in my research.

This research has been supported by Haukkalan lastenpsykiatrisen hoitolaitoksen kannatusyhdistys ry, Niilo Mäki Foundation, Vajaaliikkeisten kunto ry, Suomen Kulttuurirahaston Keski-Suomen rahasto and Jyväskylän yliopisto.

LIST OF PUBLICATIONS

1. Ahonniska, J., Ahonen, T., Aro, T., Tolvanen, A., & Lyytinen, H. (2000) Repeated assessment of the Tower of Hanoi test: Reliability, and age effects. Assessment, 7, pp. 297-310.

2. Ahonniska, J., Ahonen, T., Aro, T., & Lyytinen, H. (2000) Suggestions for revised scoring of the Tower of Hanoi test. Assessment, 7, pp. 311-320.

3. Ahonniska, J., Ahonen, T., Aro, T., Tolvanen, A., & Lyytinen, H. Practice effects of visuo-motor and problem-solving tests in children (in press)

4. Ahonniska, J., Ahonen, T., Aro, T., & Lyytinen, H. Effective or not effective? Interpreting the results of intervention in single-case research design (submitted).

CONTENTS

1 INTRODUCTION

The last two decades have seen an immense increase in research in neuropsychology and cognitive psychology. Accumulating knowledge of these research areas has lead to remarkable progress also in the field of clinical child neuropsychology. Children with learning disabilities and brain traumas are getting more accurate diagnosis, and both the professionals and the parents have better understanding of common neuropsychological disorders and their influence on daily activities.

The development of intervention methods, unfortunately, has not been able to follow the fast development in neuropsychological assessment and theory. This is partly due to logical reasons. First, efficient intervention methods are necessarily based on an accurate neuropsychological theory about normal cognitive functioning, the nature of disorders and their development. Thus, the theories of intervention can not attain a better conceptual level than the theories about neuropsychological disorders. Second, shortage of effective intervention methods results from our limited knowledge about how learning, compensation or improvement in performance takes place in each type of deficit. Third, detailed knowledge about the crucial components of intervention methods cannot be achieved unless it is possible to analyze reliably which kind of intervention influences whom as well as why and in what kind of conditions (Kazdin, 1997). Thus, designing efficient intervention methods is impossible without reliable and valid assessment of change in cognitive functions which is related to intervention.

The present thesis discusses the problems of assessing change in repeated neuropsychological assessment in children. Repeated assessment, also called time-series measurement, is used in single-case research designs in order to collect enough replications of measurement and to reach reliable conclusions. Time-series assessment, which is used in single-case research, is proving indisputable assets in clinical practice andneuropsychological research. Unfortunately, it also has several disadvantages, which are mainly related to limited generalizability of the results and to various consequences of practice effects. Both of these serious limitations of single-case research lack empirical investigation and their influence is poorly known.

This thesis consists of four publications. Studies I and II discuss in detail the effects of repeated assessment on an executive function test, the Tower of Hanoi Test (Shallice, 1982). Study I concentrates on the practice effects and reliability of various scores of the Tower of Hanoi Test, and Study II gives suggestions for revised

scoring of the test. Study III provides growth curves for four tests in two different age groups and examines how age, cognitive domain of the test, and the existence of alternative forms influence the magnitude of practice effects. Study IV presents an intervention research project, in which the single-case paradigm is applied. The data of time-series is compared with the growth curves presented in Study III, and the significance of various methods of assessing change in single-case research are discussed regarding reliable interpretation of intervention research.

2 HOW TO MEASURE CHANGE IN INTERVENTION RESEARCH

Change in neuropsychological processes is a common target of assessment in clinical neuropsychology. Change needs to be assessed when ever we follow the development of children at-risk, the progress of disease, spontaneous recovery after a trauma, or results of intervention. In intervention research, change is traditionally measured by group research design which is able to give relatively unambiguous results. The group design has a simple structure. In its simplest form, the group design consists of two homogeneous groups, out of which one group receives intervention and one not. If the improvement in performance of the intervention group is significantly larger than that of the control group, the intervention method is considered effective. In addition to statistical significance, the group design permits assessment of measurement errors and conclusions regarding generalizability of findings (Trexler & Thomas, 1992).

However, the group design also has disadvantages. Reaching statistical significance of treatment effects can be prevented by subject heterogeneity and between-subject variance. Even if statistical significance is reached, it does not necessarily equal clinical significance (Trexler & Thomas, 1992). Group design has additional limitations when applied to clinical populations. Matching the treatment groups with control groups on relevant variables in order to achieve homogeneous groups is difficult. Furthermore, it is often impossible to select participants randomly to represent any clinical subpopulation. Moreover, delay or withdrawal of intervention from part of the participants is unethical in many cases. It is also laborious to apply the group design to clinical practice. Finding a great number of suitable participants, organizing the framework, and creating smooth co-ordination between the participant and his/her family, school or work as well as coordination with the intervention team members requires plenty of planning and hard work. Last, but not least, the group design method is able to evaluate whether certain intervention is effective for certain populations, but the individual reasons for success or failure of the intervention or the timing of the improvement cannot be revealed.

Single case methodology offers a number of research designs for demonstrating change. The research designs include e.g. various forms of reversal design (ABAB), multiple baseline designs, changing criterion, and alternating treatment designs (see e.g., Hersen & Barlow,1984; Schloss, Misra, & Smith, 1992). The assessment of change is most commonly based on visual inspection by

evaluating the changes in mean, level, and trend of the data, and latency of the changes (Kazdin, 1982).

Single-case research design solves some of the methodological disadvantages of the group design. It fits easily to clinical populations, because selection of the subjects does not have to satisfy the requirement of randomization and homogeneity of the participants (Seron, 1997). Furthermore, the single-case methodology is easily applied in clinical practice because even small modifications in the routine intervention can provide enough data for research. The single-case design also provides detailed information about the reasons or timing of change (Trexler & Thomas, 1992). In cognitive neuropsychology, single-case research design is widely used for testing and verifying theoretical models of cognitive functioning (e.g. Temple, 1998)

Because the single-case research can be performed with a small number of participants, the required replications needed to show reliable changes resulting from the intervention are achieved by repeated assessments. Consequently, single-case research has several methodological disadvantages. It is prone to limited generalization of the results and to more idiosyncratic results than group design (Zurif, Gardner, & Brownell, 1989). The statistical significance of any demonstrated improvement is difficult to determine (Trexler & Thomas, 1992). Repeated assessment might also influence test-retest reliability and validity of the tests (Denckla, 1994).

Since repeated assessments with single-case research design were originally developed for assessing changes in behavior, the methodology is best suitable for behavior observation. The single-case research design also fits domains in which individual behavior is consistent, e.g. psychophysiological experiments (Caplan, 1988). However, single-case methodology is being increasingly applied in neuropsychological or psychological assessment where the dependent variable is not frequency of certain behavior, but results of various psychological or neuropsychological tests. These measures are sensitive not only to the target of research, e.g., intervention methods or progress of disease, but also to practice effects. The magnitude of practice effects are influenced by many, mostly unexamined factors: age, gender, and cognitive level of the participants, as well as contents and target area of the test, availability of alternative forms, and length of the test-retest interval. In most cases, the influence of these factors is not known, which can easily result in erroneously attributing all the changes to intervention method.

2.1 The effect of repeated assessment on practice effects

In previous experiments, the influence of repeated assessment on practice effects has been assessed by repeating the assessment once or twice after the initial assessment (see e.g., Casey, Ferguson, Kimura, & Hachinski, 1989; McCaffrey, Ortega, Orsillo & McCoy, 1992a; McCaffrey, Ortega, Orsillo, Nelles, & Haase, 1992b, McCaffrey et al., 1995; Neyens & Aldenkamp, 1996; Tuma & Appelbaum, 1980). However, in single-case research, the assessment needs to be repeated at least half a dozen times, preferably more, with relatively short intervals, in order to make reliable interpretations about the efficacy of the treatment. Thus, the previous experiments do not offer any significant

help in analyzing the amount of practice effects and interpreting data of time-series research with frequently repeated assessments. The aim of this research was to frequently repeat neuropsychological tests with relatively short test-retest intervals, and to evaluate the amount of practice effects resulting from repeated assessment.

Additionally, the aim of this research was to evaluate the effect of certain test- and person-specific variables on practice effects. In previous experiments, various neuropsychological tests have shown different amounts of practice effects. Timed tests, requiring an infrequently practiced response, or having a single easily conceptualized solution seem to result in significant practice effects (Dodrill & Troupin, 1975). Also memory tests (e.g. McCaffrey & al., 1992a; McCaffrey & al., 1992b;) and performance IQ in intelligence tests (e.g. Rawlings & Crewe, 1992; Wechsler, 1981) generally yield significant practice effects. This research provides growth curves, reliability and stability values for one visuo-motor co-ordination test, one visual discrimination test and two problem-solving tests. A special interest was focused on analyzing the influence of repeated assessment on various scores of problem-solving tests, and on analyzing whether problem-solving tests show more practice effects than visuo-motor tests.

Practice effects can sometimes be controlled by administering alternative versions of the same test in subsequent assessments. At least in word list learning (e.g. Crossen & Wiens, 1994) parallel test versions prevent practice effect. However, in the comparison between the Rey-Oesterrieth Complex Figure Test and the Taylor Complex Figure Test familiarity with the format of test may prepare the subjects for taking the slightly different test. The influence of alternative forms on practice effects is mainly unknown. In the current research, the neuropsychological tests were repeated several times using alternative forms of three tests, and it was evaluated whether the alternative forms can prevent practice effects in these tests.

Age is one of the person-specific variables, which might influence practice effects. Some studies have showed that older adults benefit less from practice than younger ones (MacNeill Horton, 1992, Mitrushina & Satz, 1991). However, there is no clear evidence which tests are most likely affected by the interaction of age and practice effects. Additionally, the interaction effect of age and practice effects is almost totally unknown in children. In this research the subjects chosen for repeated assessment were two groups of children while the mean age difference between the groups was four years. The effect of age on the amount of practice effects was assessed.

The last part of the research provides an example of intervention research, which assesses the change in the performance of two participants with pre-post measurements, behavior observations and time-series analysis. The results of the time-series analysis are compared with the growth curves provided in the previous part of the research in order to assess the efficacy of the intervention (Study IV).

3 REPEATED ASSESSMENT OF EXECUTIVE FUNCTIONS WITH THE TOWER OF HANOI TEST

3.1 The effects of repeated assessment on the results of executive function (EF) tests

From time to time EF has to be assessed repeatedly, but little is known about the effects of the repetitive assessment on reliability, or practice effects of the test. Few reports of the test-retest reliability values of the EF tests have been presented, the values ranging between .30 to .90 for the Wisconsin Card Sorting Test (WCST) (Heaton, & al.,1993; Ozonoff, 1995; Pennington, Groisser, & Welsh, 1993) between .71 to .81 for the Tower of London/Hanoi (Gnys & Willis, 1991; Culbertson & Zillmer, 1998) and above .70 for the Verbal and Figural Fluency (Gnys & Willis, 1991; Ruff, Light & Evans, 1987; Vik & Ruff, 1988). The reports of the practice effects of the EF tests are even more scarce, but the Category Test (Dodrill & Troupin, 1975) and the Trail Making Test (McCaffrey, Ortega, & Haase, 1993) have shown practice effects.

The difficulty of achieving test-retest reliability and the lack of assessing practice effects in EF tests results from the emphasis on problem solving in novel and surprising situations, a characteristic of the traditional EF tests. Unfamiliarity has been considered an essential requirement for construct validity of an EF test (Denckla, 1994; Welsh, & Pennington, 1988). If problem solving - the insight into a novel task - is considered the only essential element of EF, high test-retest reliability may not be achievable. However, PET research shows prefrontal activation occurring also in an EF task repeated many times, in cases that the task requires prefrontal activity, e.g., shifting categories in the WCST test (Weinberger, Berman, Gold, & Goldberg, 1994). It is not known how repeated assessment would influence the target ability of the assessment and the test performance. Possibly, only some parts of the EF tests could be repeated reliably (e.g. perseveration index of the WCST, Denckla referring to personal communication of Pennington, 1994).

In this research, the effects of repeated assessment on the reliability and on the practice effects of the EF tests are assessed with the TOH test. The TOH test has several advantages in a repeated assessment. First, the test can be easily modified to several alternative tests with tasks at various levels. The test also provides several different scores, such as planning time, total performance time and the number of errors which could give valuable information about EFs required for solving the test.

In most of the previous experiments only the achieved score has been counted. In the present experiment several scores of the TOH test are measured repeatedly.

3.2 Aims of the study

The first aim of this study was to assess the effect of repeated assessment on the several scores of the TOH test. The achieved score was expected to rise rapidly while the planning time, the total performance time, and the errors were expected to decrease in the first few assessments. After a few assessments, the improvement rates were expected to slow. Second, the effect of children's age on performance was analyzed. The older participants were expected to score better than the younger ones. In addition, the effect of age on the rate of performance improvement was analyzed.

A third aim of the research was to assess the reliability and the stability of the various scores of the TOH test: the achieved score, the planning time, the total performance time, and the number of errors. The reliability of any of the scores was not expected to vary remarkably with the repetitions, but the stability of the scores was expected to increase with the repetitions.

Fourth, the relation of the duration of the planning time and total performance time to the success of the performance was investigated. The successful trials were expected to have longer planning time and shorter total performance time than the unsuccessful trials.

3.3 Method

Participants

Two groups of children served as participants of the research. The younger children (n = 20, 10 boys and 10 girls) were in the first grade of school at the beginning of the study (average 7.7 years) and they finished the second grade at the end of the study. The older children (n = 28, 15 boys, 13 girls) were in the fifth grade at the beginning (average 11.6 years) and at the end of the sixth grade at the end of the research. The participants attended a normal primary school in Central Finland.

Test procedure

Because of the special requirements of repeated assessment regarding possible effects of learning, variations on the traditional TOH test procedure (Borys, Spitz, & Dorans, 1982; Welsh, Pennington, & Groisser,1991) were made in the following ways. Three alternative versions of the test were created to prevent the participants from learning the content of the tasks. The tasks did not always finish in a tower position, but flat and half tower positions were included in order to allow a greater variety of tasks. Lastly, tasks of several levels of difficulty were performed in ascending order in order to increase the sensitivity of the test.

The alternative forms of the test were created based on the state space

presented by Borys, Spitz and Dorans (1982). Three alternative versions of the TOH test were repeated three times each, resulting in nine assessments within 18 months, once every two months. The participant was introduced to the disks of different sizes as well as to two main rules: 1) a big disk is not allowed to be put over a small disk and 2) only one disk can be moved at a time. After the basic rules were explained, the children were asked to show their understanding of the rules by illustrating the incorrect moves. In the later testing sessions when the participants were more familiar with the tasks, they were asked to tell or to show only the incorrect moves.

The participants were told to move the disks on their board to exactly the same positions as on the board of the experimenter with a minimum number of moves. It was pointed out that they could use as much time as they wanted because it was not the amount of time that was important but the number of moves. The participants were also told that they could restart the particular trial from the beginning if they were not content with their current performance.

Before starting the assessment, the participants had to solve a two-disk, three-move problem in the minimum number of moves in order to be able to continue the experiment. In the test itself there were five tasks: one each of four-, five-, and six-move tasks and two seven-move tasks. The number of trials given for each task was the number of the minimum moves, less one, meaning three trials for the four-move task, four trials for the five-move task and so on. The participants had to solve each task twice consecutively in order to continue to the next, more difficult task. A participant could continue from a four-move task after only one correct solution. If the participant had not solved the task correctly on the trial before the last one, the last trial was not allowed, except in the four-move task. If the task was solved with more than the minimum number of moves allowed, the participant was told the number of moves he/she made. The participant was also requested to try to solve it again with fewer moves, but the minimum number of moves was not told. If the participant violated the rules, the trial was interrupted, the violated rule was explained, and the participant started the following trial if he/she still had an unused trial for the task in question.

Scoring

The achieved score was counted by giving, the highest score for each problem (number of minimum moves - 1) to participants who solved the problem consecutively with the minimum moves in the first and second trial, one point less if they solved the problem in the second and third trial, two points less if they solved the problem in the third and fourth trial, etc. If the participant could not complete the task even once, he/she scored zero points. If the participant could solve the task once but not twice consecutively he/she was given half a point for each isolated successful trial. Thus, the achieved score could reach a maximum of 24 points for the whole test. The total performance time used for each task was measured as well as the time used before the first move (planning time). The error score was counted as a total number of illegal moves, wrong results, and trials where the participant interrupted her/himself.

Data analysis

In order to analyze the age effect on the level and on the form of the longitudinal

profile of the TOH scores, MANOVA using profile analysis was performed. The age effect on each assessment, separately, was calculated by one-way ANOVA. The reliability and the stability of the test results were investigated by constructing a simplex model using LISREL analysis (Jöreskog & Sörbom, 1993) with the generalized least squares method (GLS). When the test results of all the participants were analyzed as one group for the LISREL analysis, the distributions of the variables were not normal, resulting in the LISREL being based on the Spearman correlation coefficients. In order to find out whether the planning time and the total performance time varied in successful vs. unsuccessful trials, the average planning time and the total performance time of each subject were calculated separately for successful and unsuccessful trials and the t-test analysis for independent groups was performed.

3.4 Results

The achieved score showed higher level of performance for the older participants than for the younger participants (see Figure 1). Also, the older participants improved their performance faster than the younger ones.

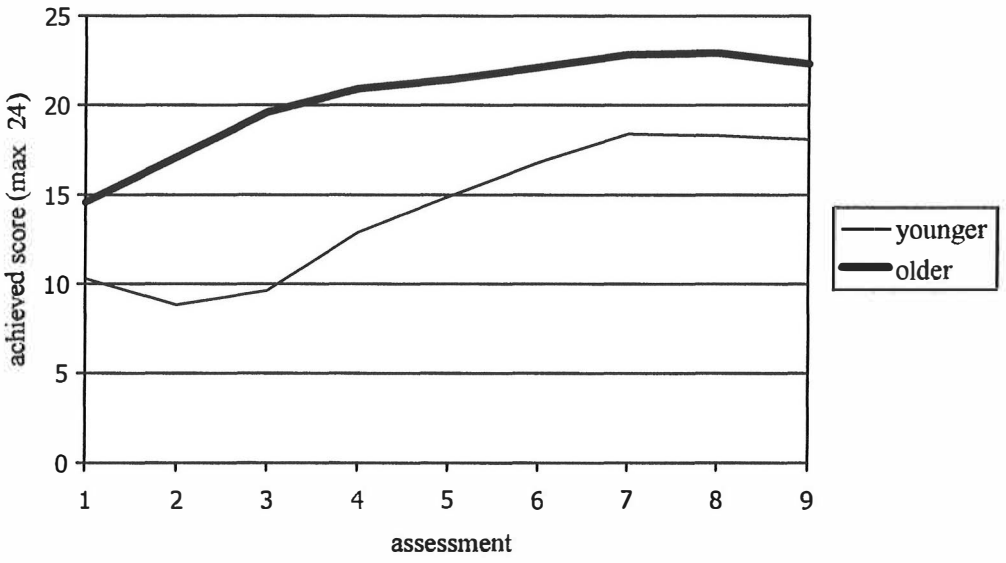


FIGURE 1 The average achieved score of the TOH test in nine consecutive assessments.

The older participants used less time for the total performance than the younger ones (see Figure 2). The amount of total performance time decreased along the repetitions in both of the groups.

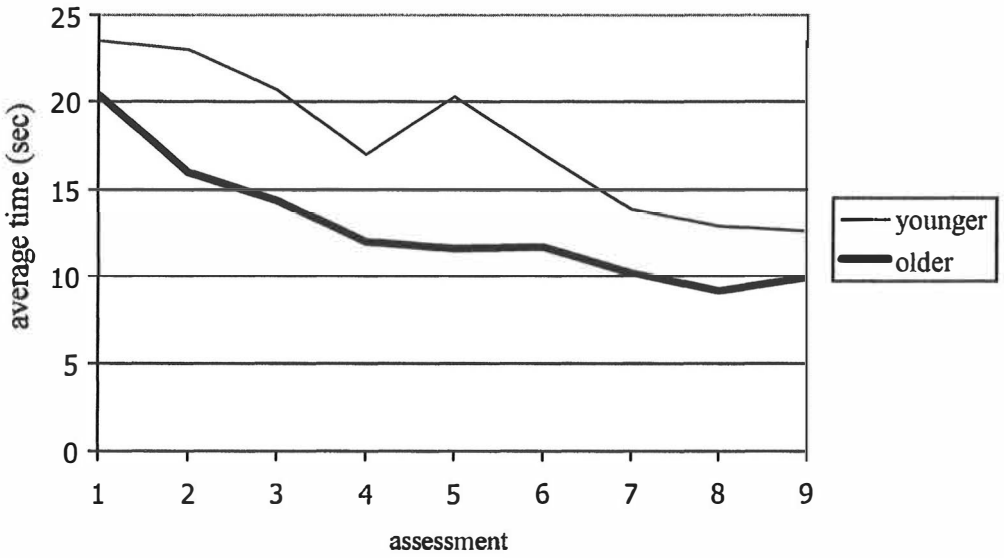


FIGURE 2 The average total performance time in the TOH test.

The amount of planning time showed a u-slope in both groups (see Figure 3), but there was no difference between the groups in the amount of time used. The amount of the errors was significantly smaller in the older group than in the younger group, but all the participants made very few errors.

The reliability of the achieved score was low in the first two assessments, but increased after the second assessment (see Table 1). The reliability of the total performance time varied between 0.6 to 1.0 in all except the sixth assessment, which was very low. The reliability of the planning time was very low in the first two assessments, but improved in later assessments. The correlation coefficients of the error score were close to zero. Thus there seemed to be a lack of reliability and stability of the error score.

TABLE 1 The reliabilities of various scores of Tower of Hanoi test

Assessment	1	2	3	4	5	6	7	8	9
Ach. score	0.48	0.49	0.63	0.67	0.54	0.54	0.58	0.65	0.67
Planning time	0.15	0.28	0.61	0.42	0.61	0.72	0.53	0.82	0.81
Total time	0.59	0.75	0.67	0.61	1.0	0.37	0.62	0.90	0.91

Note. N = 48. Values are squared multiple correlations for y-variables

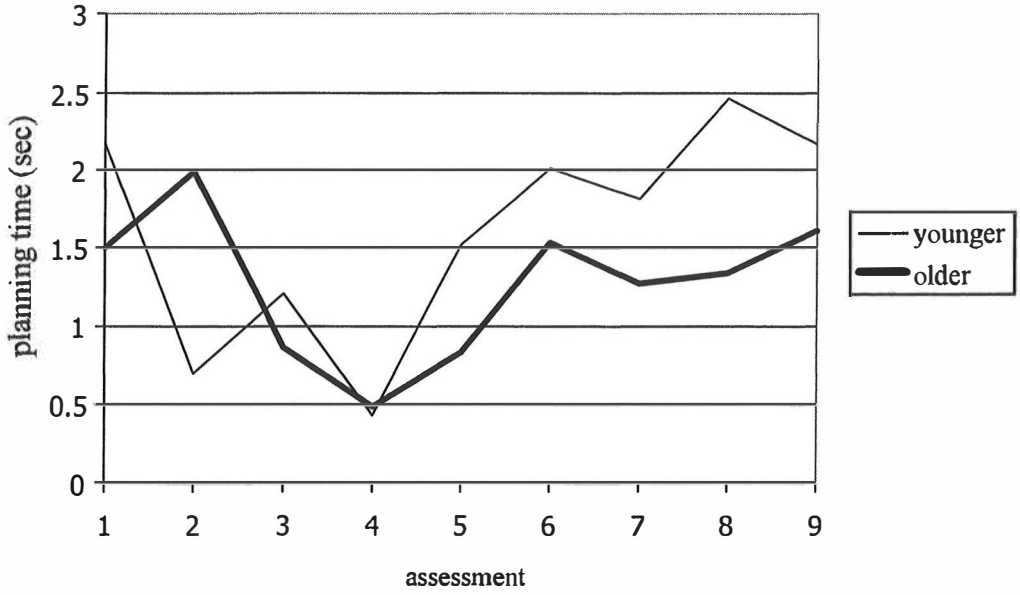


FIGURE 3 The average planning time in the TOH test.

The stability of the achieved score improved from 0.8 almost to 1.0 after the third measurement (see Table 2). The stability of the total performance time was low during the two first assessments, but improved after the second assessment. The stability of the planning time was high throughout the assessments.

TABLE 2 Stability of the scores from one measurement to another

Stability between measurements

	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9
Achieved score	0.80	0.89	0.93	1.0	0.95	0.9	0.97	0.89
planning time	1.0	0.84	0.98	1.0	0.92	0.86	0.93	0.92
total time	0.61	0.71	0.90	0.84	0.85	1.0	0.96	0.82

Note. N = 48 Values are standardized beta coefficients.

In the younger group, the planning time did not vary according to the success of the performance, but in the older group the planning time was significantly longer in the unsuccessful than in the successful trials. In both of the groups, the total performance time was significantly shorter in successful than in unsuccessful trials.

3.5 Discussion

Using three alternative forms of the TOH test did not prevent the participants from learning the general structure of the task. The achieved score improved and the total performance time decreased as a result of repeated assessment. However, the achieved score revealed an interesting difference in the learning rate between older and younger participants. The older participants already improved their performance remarkably during the first few assessments, while the younger participants started to improve their performance only after the third assessment, and even then, more slowly.

The difference between the older and younger participants in the level of performance and rate of learning, may result from the multistage development of EF occurring between these years (Passler, Isaac, & Hynd, 1985). Already at the age of six, participants show the ability to resist distractions and inhibit maladaptive responding (Becker, Isaac, & Hynd 1987; Passler, Isaac, & Hynd, 1985; Welsh, Pennington, & Groisser,1991). However, organized search, requiring greater hypothesis testing and impulse control, reaches adult levels only at the age of ten (Chelune, & Thompson, 1987; Welsh, Pennington, & Groisser,1991). Complex planning does not demonstrate adult level of mastery even at the age of twelve (Welsh, Pennington, & Groisser,1991). Additionally, the younger participants have more limited working memory in the sense of on-line mental representation of information needed for problem solution (Case, 1985). As a result, they are more impulsive and exploratory (Humphrey, 1982; Paulsen & Johnson, 1980; Vlietstra, 1982) and their selective attention in ignoring irrelevant stimuli is less developed than the older ones (Miller & Weiss, 1981; 1982). Thus, development of these various EFs improved the learning potential of the older children. Because they had learned how to learn, they were able to use the repeated assessments as learning opportunities in a more efficient way than the younger children did.

The influence of age on practice effects, visible in the achieved score, has important consequences for the applicability of the TOH test in repeated assessment. Assessing development of EF or treatment effects on EF by the TOH test should be done cautiously since practice effects could erroneously be interpreted as a real improvement of EF. In single-case studies, the recommended three measurements before starting the treatment (Barlow & Hersen, 1985) could show the practice effects in normal subjects, but achieving stable baseline could require four or five assessments before the treatment. However, less able participants could show the requested stability in the first few assessments (Pennington, Bennetto, McAleer, & Roberts, 1996) because of slow learning. In this case, the spontaneous learning occurring later could erroneously be interpreted as an effect of intervention, necessitating even larger number of assessments before starting the intervention. Thus, reliable assessment of treatment effects measured by the achieved score of the TOH test can easily become laborious and time consuming.

The reliability values of the achieved score and the planning time were low in the first few assessments, but improved from the third assessment on. Thus, in the first assessments, these scores did not measure the phenomena they were supposed to measure with satisfying reliability. This could be explained by the large intraindividual variation from one measurement to another in the first few measurement sessions as the participants were learning to solve the problem in an

efficient way. The individual learning curves revealed that most of the participants needed three assessments to learn the task and to settle their achieved score at a high level. Some of the younger participants needed even four to five assessments in order to reach an invariable, relatively high level of performance.

The standardized beta coefficient of LISREL analysis measures how stable is the ability measured by certain scores from one assessment to another. The stability values of the planning time were high already at the first assessment. The stability values of the achieved score and of the total performance time were high after the third assessment. Thus, the inter-individual differences in these scores seem to be relatively invariable already at the first assessment, becoming very constant from the third assessment on. This results at least partly from the wide age difference between the groups, which makes the performance differences relatively unchanging from the beginning.

The Spearman correlation coefficients of the error score were so close to zero in all the assessments that the reliability and stability values could not be analyzed by LISREL. This was mainly due to the low number of errors. A possible method of creating a reliable, stable and sensitive error score would be to search for milder errors and hints of erroneous thinking in addition to the ones counted in this experiment. This requires detailed analysis of the performance.

Contrary to expectations, the long planning time was not reflected as an increase in the achieved score. In fact, the older group used more planning time for the unsuccessful trials than for the successful ones. The lack of association between planning time and performance, together with the high stability of both of the time scores, could result from the invariability of the reaction style of each participant. Some children are slow or fast both at planning and at making the moves. Furthermore, effective planning is not always reflected as a long planning time (Krikorian, Bartok, & Gay, 1994; Welsh, Cicerello, Cuneo, & Brennan, 1994). Thus, the raw planning time score does not automatically measure the quality of the planning, but mainly the time invested in reacting to any stimulus. Thus, it is recommended that this score be replaced by a relative planning time score. The relative planning time score should differentiate the extra time invested in the planning of performance from the general reaction speed of the individual. Consequently, this method could show whether the theoretical association between the planning time and quality of performance can also be seen in practice.

More detailed scoring of error and time variables, as suggested above, could increase both validity and reliability of the test. Overall, improving the sensitivity, validity, and reliability of the TOH test requires detailed analysis of the performance and creation of new scores to separate plan generation from other EF which are needed in the successful performance of the TOH test.

4 SUGGESTIONS FOR REVISED SCORING OF THE TOH TEST

4.1 Time and error scores of the TOH Test

Successful solving of the Tower of Hanoi (TOH) Test requires cooperation of numerous executive functions (EFs) (see e.g., Welsh, Cicerello, Cuneo, & Brennan, 1994). However, generally only one score is counted, which is considered to measure the entirety of planning and execution of the sequence. Even if scores measuring time allocation in tasks are counted, a certain result can be obtained for several different reasons. Long planning time commonly occurs before the first move (Ahonniska, Ahonen, Aro, Tolvanen, & Lyytinen, 2000; Karat, 1982; Spitz, Minsky, & Bessellieau; 1982, Welsh, Cicerello, Cuneo, & Brennan, 1994) and theoretically indicates the existence of actual planning activity (Sternberg, 1981). However, long planning time does not necessarily separate successful and unsuccessful trials from each other (Ahonniska, Ahonen, Aro, Tolvanen, & Lyytinen, 2000; Anderson, Anderson, & Lajoie, 1996; Krikorian, Bartok, & Gay, 1994; Levin & al., 1991; Spitz, Minsky, & Bessellieau, 1985; Welsh, Cicerello, Cuneo, & Brennan, 1994). Long planning time can also indicate low quality problem solving (e.g., confusion, working memory limitations), recognition of the problem without the ability to solve it (Welsh, Cicerello, Cuneo, & Brennan, 1994), or individual slowness of responding to any stimulus (Krikorian, Bartok & Gay, 1994; Welsh, Cicerello, Cuneo & Brennan, 1994). Short planning time could indicate impulsivity and lack of planning or fast generation of good plans. Long total performance time has been associated with additional planning during execution of the sequence (Karat, 1982) and can be considered as a sign of revising a plan after an error. Nonetheless, long total performance time can also indicate slow speed of response to any stimulus (Krikorian, Bartok, & Gay, 1994) or confusion and lack of attention during the performance. In successful trials, short total performance time could indicate fast execution of the sequence after a successful plan. In unsuccessful trials, it could show ignorance of a mistake or a goal, and lack of planning altogether.

In conclusion, systematic analysis of planning time and total performance time, performed in relation to each other, as well as in relation to the achieved score, could help revealing used strategies, and reasons for failure in the task. No attempts

have been made to separate the individual variation in planning time from individual variation in general speed, which is the average moving time of the individual. Additionally, although the TOH test offers plenty of possibilities to count different errors, varying from mild to serious, reporting of errors has been rare and concentrated on only one kind of error: either the number of failed attempts (Anderson & al., 1996), or perseveration (Welsh, 1991) or illegal moves (Culbertson, & Zillmer, 1998; Karat, 1982; Klahr & Robinson, 1981; Leon-Carrion & al., 1991; Levin & al., 1996).

The work of Welsh and her colleagues (Welsh, 1991; Welsh, Cicerello, Cuneo, & Brennan, 1994) offered a very interesting beginning in the detailed analysis of the TOH test, concentrating on scoring of the error and temporal patterns. The participants seemed to divide the whole sequence into several simpler subgoals (Karat, 1982). The first moves of each of these subgoals were called critical moves (Welsh, Cicerello, Cuneo, & Brennan, 1994). The error analysis of the TOH task revealed that the participants made more errors in the critical moves than in the other moves in the sequence. The temporal analysis showed that the average pause time was longer in critical moves than in non-critical moves. Also, the good performers used more time before most of the critical moves than the poor performers used (Welsh, Cicerello, Cuneo & Brennan, 1994).

4.2 Aims of the study

In this study, a detailed error and time analysis was performed for each move. Based on the map of sequential moves presented by Borys, Spitz, and Dorans (1982), the seven move task was estimated to have two critical moves starting a sub-sequence, the first and the fifth. Increases of pause time and of errors were expected to be found at the critical moves (Welsh, 1991). The pause time before the critical moves was expected to vary together with the quality of the performance.

In order to extract the additional investment of planning from the general reaction style, relative time scores were counted by dividing the raw time scores by average move time. The relative time scores were expected to correlate better with performance than did the raw time scores. Also new error scores were counted, detecting milder errors than the traditional error score. Whereas the traditional error scores have not correlated well with the performance, the new error scores were expected to increase in sensitivity and to correlate with the performance.

4.3 Method

Participants

The error and temporal patterns of performance were analyzed in four separate videotaped assessments of eight participants. The participants were eight children (four girls and four boys, ten to thirteen years old) who were receiving rehabilitation for perceptual and problem-solving deficits. Five of the subjects had no known neurological etiology explaining their perceptual and problem solving deficits. Out

of the three with neurological etiology, one participant had cerebral palsy with spastic diplegia, one had fetal alcoholic syndrome, and one participant had a brain trauma as a result of a traffic accident, resulting in enlarged ventricles and cortical atrophy in the frontal lobes and in the right temporal lobe.

One of the children attended a normal primary school; all the others were pupils in a special school for children with neurological and motor deficits or attending a special education class of a normal primary school. Their average full scale IQ of the WISC-R test was 69, the average performance IQ 67 and the average verbal IQ 75.

Test procedure

The TOH test was administered to the participants in the context of a rehabilitation study during which the participants received treatment and were assessed repeatedly with several neuropsychological and cognitive tests. The test procedure was similar to the procedure described in study I, with the exception that all the performances were recorded by videotape.

Data analysis

The time the participants used for every single move was measured by a stopwatch. The latency of each move was measured from the time the previous move passed the top of its target peg to the time the measured move passed the top of its target peg. Because some of the tasks contained fewer than seven moves in the optimal case, the first move of the task was always counted as the first move, independent of its ordinal place in the state space. The consecutive moves were, however, counted according to their place in the map of sequential moves; for example, in the four-move task the first move was counted as first, although it was the fourth in the map of sequential moves, the second move as fifth and so on. In order to explore the variations in pause times and their relation to the performance, the average pause times of the seven first moves of the participants were calculated separately for correct trials, incorrect trials, and self-corrected trials. A trial was considered self-corrected if an erroneous move was made in the first move but corrected in the second one. Analyses were performed to determine whether the pause times of the moves varied according to the position of the move in sequence. Other analyses determined whether the length of the pause times varied according to the quality of performance.

Legitimate moves that deviated from the minimum path solution were counted as incorrect moves. The ordinal place of the incorrect moves was recorded in order to see whether the error patterns found in 15-move tasks (Welsh, Cicerello, Cuneo & Brennan, 1994) also applied to seven-move tasks. Only the place of the first incorrect move was counted, because the moves following an incorrect move would not have been in their original position.

For the Spearman correlation analysis, planning time was calculated as the pause time before the first move. Pause time before the fifth move was calculated as the second critical pause time and was counted only in the correct trials. Pause time before uncritical moves was calculated as the average pause time before the second to fourth and the sixth to seventh move. The average pause time was the sum of all the pause times (total performance time) divided by the number of moves, so that

only the optimal number of moves in each task in correct and incorrect trials were taken into account. The relative pause times before first, fifth and uncritical moves were calculated by dividing the particular pause time with the average pause time. Also, the relative pause time after the first incorrect move of a trial was counted in order to show evidence of self-monitoring after error.

In addition to the error scores calculated in the first experiment new error scores were calculated. A self-correction score was counted when a disk was moved from one peg to another immediately after it was moved the first time and when no other disk had been moved between these two moves, e.g., the small disk was moved first from peg C to peg A before moving any other disk to peg B. An almost-performed move was counted when the disk was placed on the peg or fitted onto the peg, and the participant moved it to another peg instead of releasing his/her hand from the disk. A perserative move was counted when the participant repeated at the beginning of the trial the same incorrect sequence she/he had performed in the immediately preceding trial. The sequence included at least two moves. The perseverations were counted from the second perseverated trial on. Two cumulative error scores were calculated: serious errors in self monitoring (illegal moves, wrong results and perseverative moves) and mild errors in self monitoring (almost performed moves, interrupted trials and self-corrected moves). The average number of moves in an incorrect trial was measured and an especially large number of moves were considered to show a trial and error style.

4.4 Results

The average raw pause times before every move are shown separately in differently performed trials in Figure 4. The repeated measures ANOVAs with Greenhouse-Geisser correction revealed that there were differences between the pause times in all the various trials: correct trials, incorrect trials and self-corrected trials. In the correct trials, the pairwise comparison of the pause times revealed the pause time before the first move to be significantly longer than any other pause time (see Table 3). The pause time before the fifth move was significantly longer than before the non-critical moves, except for the third move. In the incorrect trials, the pause times before the first, second, and third moves, were significantly longer than the pause times before fifth, sixth and seventh move. In the self-corrected trials, the pause time was longest before the second move and second longest before the first move. The rest of the pause times did not differ from each other.

The relation of performance to pause times was analyzed by repeated measures ANOVAs with Greenhouse-Geisser correction (see Figure 4). The pairwise comparison revealed the pause time before the first move to be shorter in the self corrected trials than in the correct and incorrect trials. All the other pause times were significantly shorter in the correct trials than in the incorrect trials, aside from the pause time before the fifth move. The second pause time of the self-corrected trials was longer than the second pause time in the correct trials. The third, the sixth and the seventh pause times were shorter in the self-corrected than in the incorrect trials.

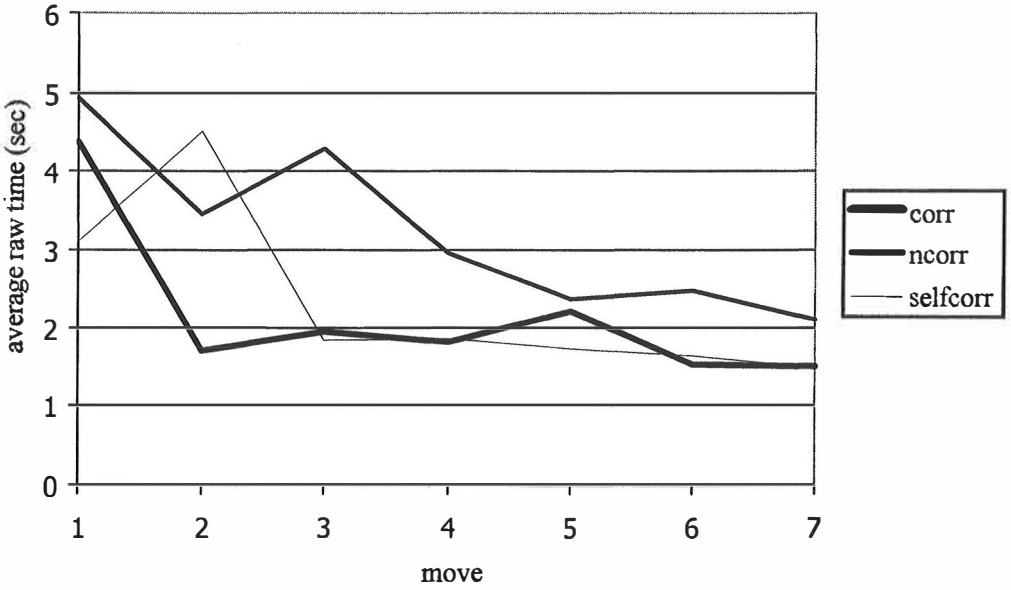


FIGURE 4 The average pause time used for seven first moves in correct trials, incorrect trials and the trials where the first move error was corrected immediately (N = 8, four assessments for each).

When the relation of raw time scores, relative time scores, and errors to performance was examined by the Spearman correlation analysis, the raw planning time had a positive although relatively weak correlation with the achieved score (.36). The fifth move time (-.26) did not correlate significantly with the achieved score. Neither did the relative fifth move time (.09). However, the relative planning time correlated positively and statistically very significantly with the achieved score

TABLE 3 The ANOVA pairwise analysis between the Tower of Hanoi average pause times of moves

Move	1	2	3	4	5	6	7
				correct trials			
1st	--	82.75***	57.99***	88.73***	53.38***	113.77***	109.00***
2nd			4.93	0.10	14.01**	2.56	2.80
3rd				0.55	3.25	4.46	4.41
4th					35.44**	20.51**	27.23**
5th						71.90***	55.97***
th							0.20***

(Table 3 continues)

(Table 3 continues)

incorrect trials

1st	--	5.37	0.60	10.21*	22.79**	20.62**	35.66**
2nd			3.76	1.67	5.65*	10.52**	8.22*
3rd				3.32	6.30*	9.71*	7.07*
4th					1.98	1.10	5.16
5th						0.17	2.12
6th							1.35

self-corrected trials

1st	--	6.59*	9.45*	9.09*	5.28	8.28*	8.76*
2nd			13.95*	18.39**	11.20*	15.81*	82.57***
3rd				0.03	0.27	0.47	2.67
4th					0.33	1.65	2.90
5th						0.05	4.70
6th							0.33

Note. N = 8.The values are F- values
*p < .05 **p < .01 ***p < .001

(.61), negatively with the serious errors (-.45) and perserevative errors (-.43). The average move time correlated negatively with the achieved score (-.64). The sum of serious errors correlated negatively with the achieved score (-.39) and positively with the average move time (.42). The sum score of mild errors did not correlate significantly with the achieved score (-.07). The relative pause time after error did reach a relatively high, although not a significant correlation with the achieved score (.30).

4.5 Discussion

The results of detailed analysis of the TOH performance indicate that specific time and error patterns could reflect the existence or absence of problem solving strategies. The time analysis revealed that the participants used more planning time before critical moves than before non-critical moves. Also, an increase of errors was found at the first move but not at the fifth move. The raw planning time correlated positively, but relatively weakly, with the achieved score. However, the relative planning time correlated positively and very strongly with the achieved score, and negatively with the serious errors. The serious errors correlated negatively with the achieved score, but the milder errors or any individual error scores did not.

As expected, the length of the pause time was associated with the quality of the performance. First, the raw pause time of correct trials showed disproportional time investment before critical moves (first and fifth). Second, in the self-corrected trials, the pause time before the second, error-correcting move was longer than before any other move, thereby reflecting the extra time invested in planning and its positive

effect. In the self-corrected trials also, the pause time before the first move, was significantly shorter than in the correct and incorrect trials. This could be an indication of impulsivity, since the capacity to solve the problem obviously existed. Third, the relatively long pause times all along the sequence of incorrect trials, reflect later attempts to revise plans in order to solve the problem.

Nonetheless, the raw pause time before the first move and the quality of performance were not directly related. The planning time was long both in correct and incorrect trials, even slightly longer in the incorrect ones. The correlation analysis shed more light on the contradictory relations between planning time and quality of performance. The relative planning time correlated with the performance, but the raw planning time did not. There also was a negative correlation between the relative planning time and serious errors. Thus, the relative planning time seems to extract the extra time invested for successful planning, from the general reaction speed, confusion, and inattention.

Out of the several error scores, only the sum score of serious errors correlated negatively with the performance. The small number of each individual error type might account for their lack of correlation with the performance. Thus, if lack of self-monitoring is operationalized by the error scores, one needs to count several types of errors. The lack of negative correlation between the mild errors and the performance might result from the fact that the mild errors (almost performed moves, self corrected moves and interrupted trials) measure more the revision of plans and slightly delayed self monitoring, than total lack of planning.

The TOH test is a promising tool for assessing complex problem solving abilities. The detailed time and error analysis could enable assessing several components of EF, instead of counting only one score requiring the efficiency of numerous EF. On the basis of this experiment, following subprocesses could be suggested for further analysis: 1) quality of planning (relative planning time, in 15-move tasks also relative critical move time) 2) execution of the planned sequence (achieved score) 3) lack of monitoring one's own behavior (sum of illegal moves, wrong results and perseverations) 4) revising plans when needed (amount of time used after incorrect move, sum of interrupted trials, self corrected moves, and almost performed moves).

In order to assess the relevance of these findings, time and error analysis needs to be done with larger samples of normal participants. Also comparisons between normal participants and participants with EF deficits are needed. The validity analysis of suggested scores would need also comparison to other EF tests measuring planning, impulsivity, perseveration, rule breaking, and plan revision.

5 PRACTICE EFFECTS IN VISUO-MOTOR AND PROBLEM SOLVING TASKS

5.1 Practice effects in repeated neuropsychological assessment

Repeated assessment commonly results in improvement of the results even without intervention (see e.g. McCaffrey, Ortega, Orsillo, Nelles, & Haase, 1992b). Thus, in order to make the correct conclusions in time-series measurement, it is very important to differentiate development or intervention effect from practice effects. Practice effects are influenced by several person- or test-spesific factors. Among these the length of the test-retest interval is very crucial. If the test-retest interval is one year or more, no practice effects have been found (Dikmen, Machamer, Temkin, & McLean, 1990; Uchiyama & al., 1994). When a test-retest interval varies between one week and six months, significant practice effects are shown in a wide selection of psychological tests (see e.g. McCaffrey, Ortega, Orsillo, Haase & McCoy, 1992a; McCaffrey, et al., 1992b, McCaffrey & al., 1995; Neyens & Aldenkamp, 1996). The shorter the test-retest interval, the stronger the practice effect (Catron & Thompson, 1979; Schuerger & Witt, 1989).

Various tests show a different amount of practice effects. Timed tests, requiring an infrequently practiced response (the Trail Making Test, the Visual Search Test, the Paced Auditory Serial Addition Task), or having a single, easily conceptualized solution (e.g., the Category Test), show large practice effects (Dodrill & Troupin, 1975; McCaffrey, et al., 1992a, McCaffrey, et al., 1992b; McCaffrey, Ortega, & Haase, 1993; McCaffrey, et al., 1995). Also, commonly used memory tests yield significant practice effects (McCaffrey, Ortega, Orsillo, Nelles, & Haase, 1992b, McCaffrey, & al., 1995). However, using parallel test versions in word list learning prevents practice effects (Crossen & Wiens, 1994; Parker, Eaton, Whipple, Heseltine, & Bridge, 1995). In the WAIS-R test, the performance IQ shows a greater practice effect than the verbal IQ (Catron & Thompson, 1979; Dodrill & Troupin, 1975; Rapport, Brooke-Brines, Axelrod, & Theisen, 1997; Rawlings & Crewe, 1992; Wechsler, 1981). Consistent performances without practice effects have been found on the tests measuring auditory or sensory perception, motor steadiness, reaction time, or focused alertness on a task (McCaffrey, Ortega, & Haase, 1993; McCaffrey et al., 1992a; McCaffrey, et al., 1992b).

Children show practice effects in the intelligence tests in a similar way to adults. In the Wechsler Intelligence Scale for Chilren – Revised (WISC-R) and in the

Wechsler Preschool and Primary Scale for Children (WPPSI) tests, the performance IQ showed larger practice effects than the verbal IQ at a test-retest interval of six months (Neyens & Aldenkamp, 1996; Tuma & Appelbaum, 1980). At the same interval The Developmental Test of Visual- Motor Integration (VMI) did not show any significant practice effects, while the Children's Paced Auditory Serial Addition Test (CHIPASAT; Dyche & Johnson, 1991), the Stroop test, the Word Test (Dutch memory test), the Rey Auditory Verbal Learning Test, the Rey Complex Figure Task and part B of the Trail Making Test show significant practice effects (Neyens & Aldenkamp, 1996).

Age and cognitive level influence practice effects. Among adult participants, the young and middle-aged show larger practice effects than the elderly ones (MacNeill Horton, 1992; Mitrushina & Satz, 1991). Additionally, high intelligence increases practice effects (Rapport, Brooke-Brines, Axelrod, & Theisen, 1997). There are no studies reporting the influence of age on practice effects of children, but it could be hypothesized that the older subjects show larger practice effects than the younger subjects, because of their higher cognitive capacity. In children, development complicates the interpretations of practice effects. Development effect might account for improved performance at a test-retest interval as brief as six months (Levin, Ewing-Cobb, & Fletcher, 1989).

Using parallel forms of the existing tests might prevent practice effects at least in memory tests (Crossen & Wiens, 1994; Parker & al., 1995). In problem solving tasks, repeating the same task format in an alternative task would presumably create practice effects (Denckla, 1994), although practice effects might be smaller than those obtained using the same task (McCaffrey, & al. 1992b). Up until now, very little has been known about the influence of alternative forms on practice effects.

Previous experiments have repeated the same test only two or three times in order to examine practice effects. In single-case studies of intervention research, the abilities of the participants need to be assessed more frequently with the test-retest interval of several weeks. Thus, growth curves are needed to differentiate intervention effects from practice effects (Denckla, 1994).

5.2 Aims of the study

In this research, a test-retest interval of two months was used. This experiment provides growth curves, reliability and stability values of one visuo-motor coordination test, two problem solving tests, and one timed visual discrimination task. Three questions are addressed. First, would alternative forms prevent or diminish the practice effect in problem solving tasks? Second, if alternative versions do not prevent practice effects, do various tests show a different amount of practice effect? One could hypothesize that the visuo-motor task would show less practice effect than the problem solving tasks, even if alternative forms for the problem solving tasks were used. Third, how does age influence the practice effect among children? It could be expected that the better cognitive abilities of older children result in larger practice effects in the older participants than in the younger ones.

5.3 Method

Participants and the procedure

The participants were the same as described in the study I (the younger group 7.7 years and the older group 11.6 years in the beginning of the study). The average score of the Colored Progressive Matrices (CPT; Raven, 1965) was 24.1 for the younger group (*z*-value 0.63, compared to the local normative sample, Niilo Mäki Institute, 1992), and 34.4 (*z*-value 0.80) for the older group. The standard scores were not significantly different in the two groups.

Measures

The Developmental Test of Visuo-Motor Integration (VMI). The VMI test (Beery, 1982; 1989) is the most widely used developmental test of visuo-motor coordination. The test has 24 geometrical figures presented in ascending order of difficulty and the possible range of the scores is from 0 to 50.

The Underlining Test (UL). The UL test (Doehring, 1968, Rourke & Gates, 1980; Rourke & Petrauskas, 1977) assesses speed and accuracy of visual discrimination. In this research, three subtests of the UL test were used. They were subtest 1 (Single Number), subtest 7 (Two Letters), and subtest 8 (Sequence of Geometric Forms). Out of each subtest three alternative versions were created by changing the letter, number or the content of geometric sequence to be searched and its location in the row of stimulus. The dependent variable consisted of the cumulative net score of all the three subtests (correct items minus errors). The maximum score was 117.

The Porteus Mazes Test (PMT). The PMT (Porteus, 1965) consists of a series of increasingly difficult mazes which are designed to measure successful planning, inhibition of impulses, and ability to change set. The PMT has three versions (Vineland, Extension, and Advanced) which are not parallel. The Extension version is more difficult than the Vineland, but easier than the Advanced version. The Vineland version was used at the first, fourth and seventh assessments, the Extension at the second, fifth and eighth, and the Advanced at the third, sixth and ninth assessments. The minimum score was 7 and the maximum 17 for each of the versions.

The Tower of Hanoi Test (TOH). The TOH test is a disk-transfer task (e.g., Borys, Spitz, & Dorans, 1982; Klahr, & Robinson, 1981; Shallice, 1982; Simon, 1975), which evaluates several areas of executive functioning. The highest score possible was 24 points. A detailed description of the TOH test and the assessment procedure can be read in chapter 3. The VMI test and the UL test were administered in a classroom setting, the TOH test, and the PMT individually.

Data analysis

Age, development, and practice effects could have influenced the results of the repeated assessments. The age difference between the two participant groups was four years (48 months). The effect of age on the results of each measurement (interval two months) was estimated by calculating the difference in score between

the older participants and the younger participants at the first assessment. This score was then divided by 24, and the result was an estimated development effect of two months. Starting from the second assessment, the calculated development effect of each test was then subtracted from each assessment result; that is, second assessment minus two months' age effect, third assessment minus four months' age effect, etc. After this subtraction, significances of the practice effects were calculated by the repeated Manova using Greenhouse-Geisser correction; 9(repetition) x 2(age) x 2(gender).

Reliability and stability of the tests were investigated by constructing a simplex model using LISREL analysis (Jöreskog & Sörbom, 1993) with the generalized least squares method (GLS). When the test results of all the participants were analyzed as one group for the LISREL analysis, distributions of the variables were not normal, so LISREL was based on the Spearman correlation coefficients.

5.4 Results

The VMI test did show main effect of age, but no main effect of repetition. The interaction effect of repetition and age was not significant showing no difference in practice effects between the groups (see Figure 5).

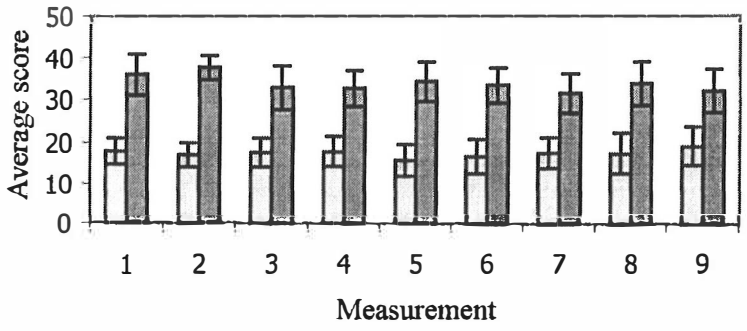


FIGURE 5 The development corrected results of repeated assessment of the Developmental Test of Visuo-Motor Integration in the older (n = 28, darker column) and in the younger (n = lighter column) group. Maximum value 50. Standard deviation displayed for each measurement.

The main effect of repetition in the UL test was significant, as well as the main effect of age. The practice effects were significant both in the younger and in the older group, and the interaction effect of repetition and age was also significant, showing that the older group showed larger practice effects than the younger group (see Figure 6).

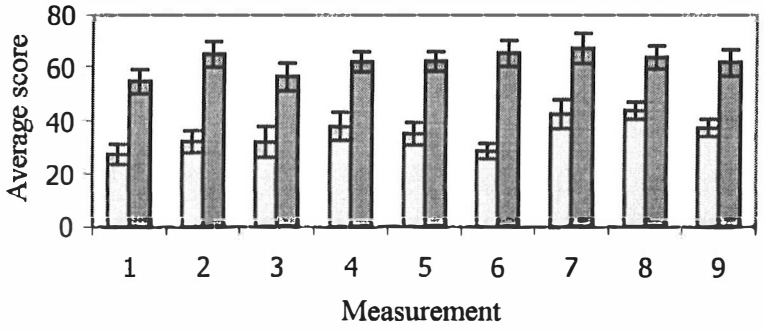


FIGURE 6 The development corrected results of repeated assessment of the Underlining Test in the older (n = 28, darker column) and in the younger (n = 22, lighter column) group. Maximum value 117. Standard deviation displayed for each measurement.

Also in the PMT the main effect of repetition and age was significant as well as the interaction effect of repetition and age (see Figure 7), demonstrating the larger practice effects in the older than in the younger group. In both of the groups, the PMT showed significant practice effects.

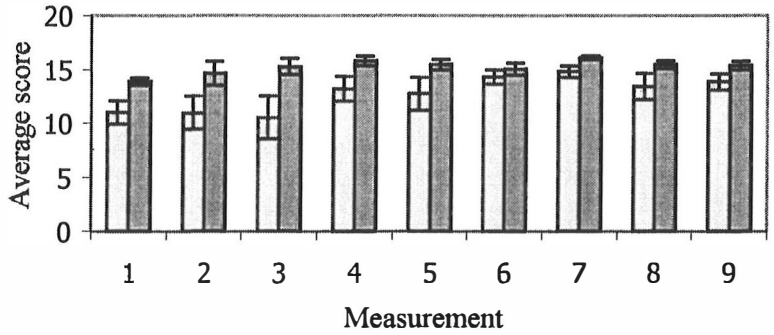


FIGURE 7 The development corrected results of repeated assessment of the Porteus Mazes Test in the older (N = 28, darker column) and in the younger (N = 22, lighter column) group. Maximum value 17. Standard Deviation displayed for each measurement (Insert here Figure 8)

The TOH test showed significant main effects of repetition and age. Significant practice effects were found in the younger group and in the older group (see Figure 8). Significant interaction effects of repetition and age showed larger practice effects in the older than in the younger group. Any of the aforementioned tests did not show significant main or interaction effects for the gender variable.

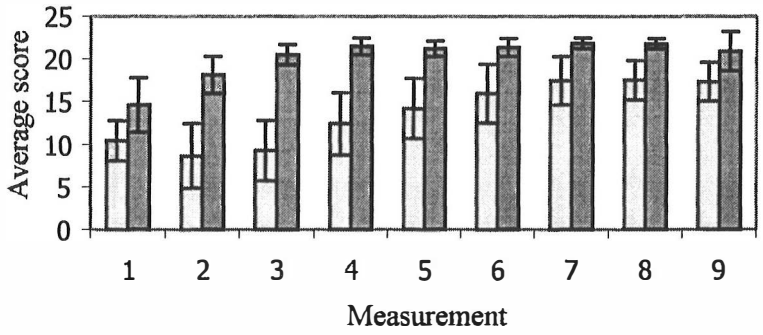


FIGURE 8 The development corrected results of repeated assessment of the Tower of Hanoi Test in the older (n = 28, darker column) and in the younger (n = 22, lighter column) group. Maximum value 24. Standard deviation displayed for each measurement.

The reliability values of the VMI test and the UL test, measured by the squared multiple correlations for y-variable, were high all throughout the assessments (see Table 4). The reliability values of the PMT and the TOH test were lower than those of the VMI test and the UL test. The reliability values of the PMT decreased slightly during the assessments while the reliability values of the TOH test increased after the first two assessments.

TABLE 4 The reliabilities of various neuropsychological tests (Squared Multiple Correlations for y-variables), n = 48.

	Measurement								
Test	1	2	3	4	5	6	7	8	9
The Developmental Test of Visuo-Motor Intergration	0.81	0.81	0.81	0.81	0.81	0.80	0.80	0.80	0.80
The Underlining Test	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
The Tower of Hanoi	0.48	0.49	0.63	0.67	0.54	0.54	0.58	0.65	0.67
The Porteus Mazes	0.65	0.64	0.64	0.63	0.62	0.61	0.61	0.60	0.59

The stability of the tests was described by the standardized beta coefficient (see Table 5). The stability of all the tests, except for the TOH test, was high throughout the assessments. The test results of the TOH started to be relatively stable from the third assessment on.

TABLE 5 Stability of the tests from one measurement to another (standardized beta coefficient)

	Measurements compared							
Test	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9
The Developmental Test of Visuo-Motor Intergration	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
The Underlining Test	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
The Tower of Hanoi	0.80	0.89	0.93	1.0	0.95	0.90	0.97	0.89
The Porteus Mazes	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

5.5 Discussion

This research provides growth curves of one visuo-motor test, one visual discrimination test and two problem solving tests. The tests showed different amounts of practice effects. No practice effect was found in the VMI test, whereas the UL test, the PM test, and the TOH test showed significant practice effects in both of the groups. Even alternative versions of the same test did not prevent significant practice effects either in the problem solving tasks or in the visual discrimination task. Probably, in repeated assessment, the participants learn the basic rules, task-appropriate strategies and plan generation; therefore changes in the task content can not prevent improvement in the performance. The older participants showed larger practice effects than the younger ones in all the tests, except for the VMI test. The reliability and stability values of the VMI test and the UL test were high throughout the assessments, whereas the reliability values of the TOH test and of the PM test were lower. The results of the PM test were very stable, but the results of the TOH test started to be relatively stable only after the first two assessments.

In this experiment, the VMI test was administered without alternative versions, while the UL test and the TOH test had alternative versions which were equally difficult, and the PMT had three versions of increasing difficulty. The three increasingly difficult versions of the PMT are supposed to prevent practice effects. The current results show that every three consecutive assessments of the PMT were on the same level, this was followed by a rise in the curve when the easiest version was administered after the most difficult one. Thus, the alternative versions of the PMT are able to prevent practice effects. The increasingly difficult versions of PMT provides another model in preventing practice effects in repeated assessment. This method might also be applicable to other tests and help to interpret the results of intervention research.

It could be argued that our method of estimating the development effect results in bias due to the non-linear development between the ages of seven and nine, and eleven and thirteen. Development in the older group could be remarkably slower than in the younger group, resulting in underestimating the development effect in the younger group and overestimating it in the older group. However, when the current

method of calculating the practice effect was compared to the available developmental data on the VMI test (Beery, 1989), the PMT (Krikorian, & Bartok, 1998), and the UL test (Rourke, & Gates, 1980), the standardized scores gave results similar to our method. In both of the groups, the standard score of the VMI did not change from the first to the last assessment in both of the groups. In the UL, the average score improved by one standard deviation from the first to the last assessment in all the subtests of the younger group and in two out of three subtests in the older group. In the Vineland version of the PMT, both groups improved their average score by one standard deviation from the first to the seventh assessment. Thus, the method of calculating practice effects in the current experiment provides reasonable results at least in the three tests with the available standard scores.

The problem solving tasks, in addition to being sensitive to practice effects were less reliable measures than the VMI test and the UL test. The individual learning curves showed great variability in the results of the TOH test and the PMT in the first few assessments. Most of the subjects needed three measurements to stabilize their results at a relatively high level. The decrease of reliability values of the PMT in the last assessments probably results from the ceiling effect.

The stability values of the VMI test, the UL test and the PMT showed that inter-individual differences were very constant from the first to the last assessment. This is probably explained by the wide age difference between the groups. The stability of the TOH test was not satisfactory in the first two measurement sessions, but improved after the third assessment. However, the inter-individual differences in the TOH test did not ever become as stable as in the other tests.

As a summary, it could be assumed that if the score of the VMI test shows significant improvement in intervention research, the intervention very likely improves visuo-motor coordination. Significant rises in the three other tests could also result from practice effects.

6 CLINICAL DECISION MAKING IN SINGLE-CASE INTERVENTION STUDY

6.1 Demonstrating intervention effects in group and single-case designs

Neuropsychological rehabilitation research has two major goals: to develop a variety of effective intervention methods, and to understand how, why, for whom and in which conditions the particular intervention methods are effective. Without sensitive, valid and reliable assessment of change, the efficacy of the intervention methods can not be evaluated. Traditionally, demonstration of intervention effects is based on group studies. The group design, although it has its assets with normal population, has several disadvantages when applied to clinical population. Appropriate selection of the participants, matching and randomizing them in order to generate homogeneous groups is seldom possible (e.g., Seron, 1997). Additionally, dividing participants into intervention and control groups may create ethical problems in clinical settings. Furthermore, applying group research design to clinical work is laborious and the data is rarely achieved as a side product of a clinical enterprise, while routinely performed clinical intervention can provide enough data for a single-case research. From a cognitive perspective it can not be assumed that all patients with similar symptoms would be suffering from same deficits or would be similarly influenced by the treatment (Caramazza, 1986). Thus, if the intervention research wants to explain which intervention works for whom and why, single-case design is commonly preferable (Kazdin, 1997).

The commonly used dependent variable for time-series analysis is based on behavior observation. Nevertheless, a great amount of theoretically and clinically relevant change resulting from neuropsychological intervention, can not be assessed by behavior observation, but requires other measures such as intervention related tasks, neuropsychological or achievement tests. However, most of the neuropsychological and achievement tests are not suitable for repeated assessment because they do not have parallel or alternative forms. Additionally, practice effects have been demonstrated in some neuropsychological tests even if the test has alternative forms, i.e., repeating the same structure with different contents does not prevent practice effects (Ahonniska, Ahonen, Aro, Tolvanen, & Lyytinen, in press). Various tests have also demonstrated different amounts of practice effects

(Ahonniska & al., in press). In order to conclude whether the improvement of the performance reflected in time-series data demonstrates genuine intervention effect or is the result of practice effects, the data of the time-series measurement should be compared to a growth curve of the same tests carried out without intervention.

6.2 Aims of the study

In this study, we presented and compared different methods for demonstrating intervention effects in single-case research. The data includes pre-post measurements, behavior questionnaires filled out by parents and teachers during the intervention, time-series measurement based on a multiple baseline across participants design, and comparison of time-series measurement to the growth curve of a control group receiving no intervention. Additionally, we wanted to assess procedure-specificity and generalization of treatment effects by measuring performance in tasks related to intervention material with varying degrees. Procedure-specificity of intervention method indicates whether change in performance is restricted to the material or to cognitive functions trained in the intervention (Seron, 1997). Generalization of treatment effects investigates whether there occurs transfer of treatment effects from one training session to another or to alternative forms of training material (lowest level of generalization), to tests that measure cognitive functions which are closely related to the intervention method (second level of generalization), to day-to-day functioning (third level of generalization; see e.g., Gordon, 1987).

In this research, the time-series data of the intervention-related tasks reveals whether the subjects improved their performance in the tasks similar to the intervention material (no-transfer). The time-series data of tasks closely related to the target area of intervention (visuo-perceptual functioning, problem solving), reveals the amount of near-transfer, and the pre-post measurement and behavior evaluations serve in assessment of far-transfer effects of intervention. The procedure-specificity of the intervention was assessed by comparing the results of no-transfer tasks with tasks, which were not included to the intervention material.

6.3. Method

Participants

The participants were two girls, Mira, 9.1 years old at the beginning of the research, and Anni, 9.2 years old at the beginning of the research. Mira had a right hemisphere brain injury resulting from pre-term birth, and Anni had a fetal alcohol syndrome. Both of the participants had deficits in visuo-motor and problem solving functions as can be seen in the results of the neuropsychological and intelligence tests of Figure 9. Their language functions were relatively well developed and both of them were able to read fluently. They attended a special school for children with

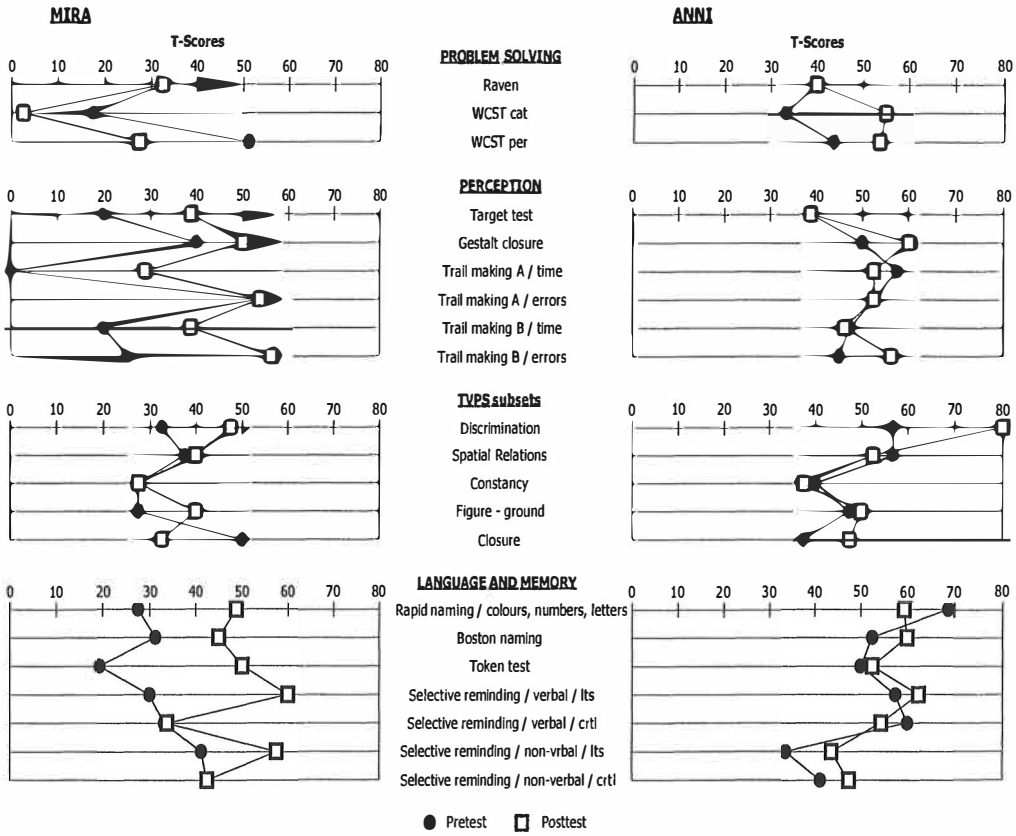


FIGURE 1 The results of the pre-post measurement of the neuropsychological tests. The tests: The Progressive Matrices (CPT, Raven, 1965), the Wisconsin Card Sorting Test/categories and perseveration (Heaton, Chelune, Talley, Kay, & Curtiss, 1981), the Target Test (Reitan & Davidson, 1974), the Gestalt Closure (Kaufman, 1983), the Trail Making Test (TMT, Reitan & Wolfson, 1992), the Test of Visual Perceptual Skills/subtests: visual discrimination, spatial relations, form constancy, figure-ground discrimination, visual closure , (TVPS, Gardner, 1982), the Rapid Naming Test (Denckla & Rudel, 1976), the Boston Naming Test (Kaplan, Goodglass & Weintraub, 1978), the Token Test (DeRenzi & Vignolo, 1962), the Selective verbal memory (Buscke & Fuld, 1974), the Selective non-verbal memory (Fletcher, 1985).

neurological and motor disorders. The IQ of Mira was 63 and the IQ of Anni was 86, measured by the Wechsler Intelligence Scale for Children- Revised (1974).

The control group (n = 22) for time-series intervention was not compared with the participants according to their age (average age 7.7 years), but according to their performance in the tests of Coloured Progressive Matrices (CPT, Raven, 1965), and the Underlining test (UL, Rourke & Gates, 1980). At the beginning of the research, the average score of CPT of the control group was 24.1 (SD 6.9). The CPT result of Anni was 23 points, and the score of Mira 16 points. The UL result of the control group was 29.8 (6.5), of Anni 36 points and of Mira 18 points. The participants of the control group were pupils of a normal primary school in Central Finland.

Intervention method

The Instrumental Enrichment method (Feuerstein, 1980) was used as a framework for the intervention. The intervention material included the nine pages of Dots and the whole workbook of Orientation in Space I, as well as 3-5 pages of both Illustrations and Comparisons. Intervention was given twice a week for 45 minutes in a group of two children with one psychologist as the therapist. The treatment period was 12 months (two and a half semesters) and the total amount of intervention was 60 hours. The aim of the intervention was to improve visuo-perceptual and visuo-motor functioning of the participants as well as executive functions (e.g., concentration on the task, planning, inhibition of impulsivity, evaluation of the performance).

Measures

Pre-post measurement included a variety of neuropsychological tests, see Figure 9. The pretest was performed one month before the beginning of the intervention and the posttest, one month after the end of the intervention.

Time-series measurements were performed every two months. Mira had three and Anni five baseline measurements before the intervention. Altogether, Mira was assessed nine times and Anni twelve times.

Time-series analysis included following tests:
The Tower of Hanoi (TOH, Shallice, 1982; Borys et al., 1982 Welsh et al., 1991), the Porteus Mazes Test (PMT, Porteus, 1965), the Developmental Test of Visual-Motor Integration (VMI, Beery, 1989), and the Underlining Test (UL, Rourke & Gates, 1980). All these tests were supposed to measure near-transfer effects of the intervention. The contents and administration of these tests have been explained in more detail in chapter 5. Additionally, three tests, the Dots, the Orientation, and the Comparisons, were created on the basis of the intervention material and used in the time-series measurement. The Dots and the Orientation were tasks requiring no-transfer. They were based on the intervention material and were used for assessing whether the participants learned to solve tasks directly related to the intervention material. The maximum score of the Dots test was 26 and the maximum score of the Orientation test was 12. The material for the Comparison test was taken from the intervention material, which was not used in this intervention. The task was used to assess whether improvement of performance was restricted to the intervention-related tasks or whether it was visible in unrelated tasks as well. The maximum score of this task was 21.

Teachers and parents filled out *questionnaires* evaluating the behavior of the participants once each semester, that is, once in four months. Altogether, four behavior evaluations were filled out for Mira and five for Anni. Mira had one and Anni two behavior evaluations filled before the beginning of the treatment.

Parents and teachers filled out the Teacher's Self-Control Rating Scale (Humphrey, 1982). A total sum score was calculated for the analysis. Minimum score was 15, and maximum score was 75, indicating good cognitive and behavioral self-control.

Additionally, the teachers filled out the Academic Performance Rating Scale (Barkley, 1991) and the Deveraux Elementary School Behavior Rating Scale (Spivack & Swift, 1982). The total scale of the Academic Performance Rating Scale of Barkley was used. The total scale includes the scales of Learning Ability, Impulse Control, Academic Performance, and Social Withdrawal. The range of the values

could vary between 19 and 95, the maximum score reflecting academically successful behavior.

The Deveraux Elementary School Behavior Rating Scale does not have a total score, but has several subscales. In the current experiment, scales of peer co-operation (range of values 2-14), work organization (4-26), perseverance (2-12), impatience (4-22), irrelevant thinking (4-20), and inattention (4-24) were used in order to assess changes of behavior during the intervention. The maximum of the first three scales showed desirable behavior, while the maximum of the three last scales showed undesirable behavior. All the behavior questionnaires as well as the neuropsychological pre-post measurement were used to assess far-transfer effects of the intervention.

6.4 Results

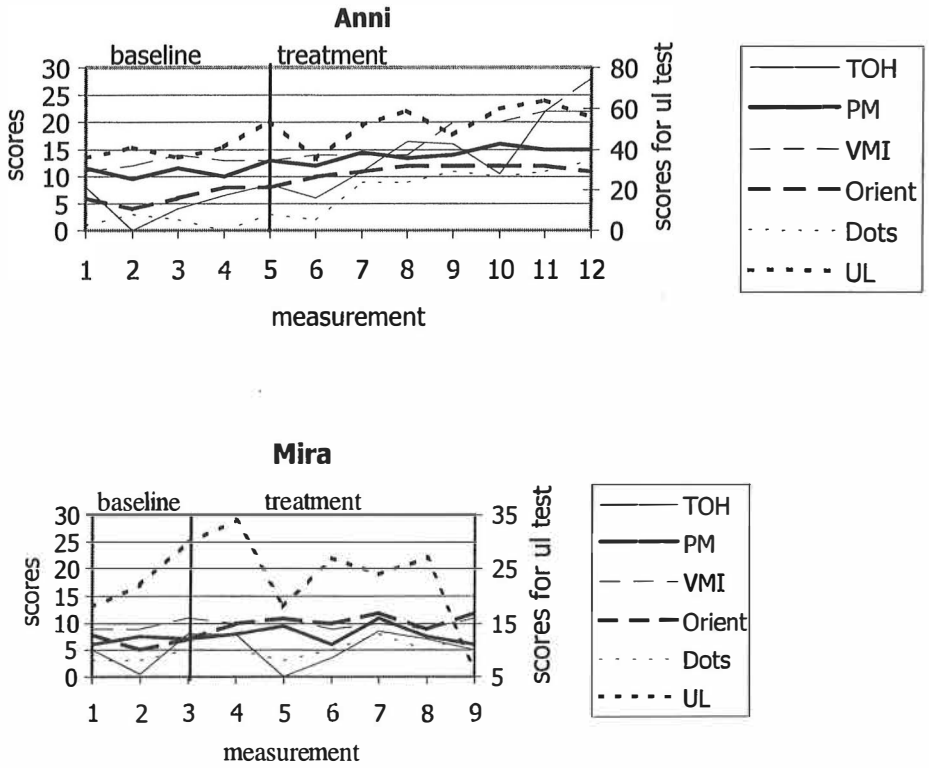


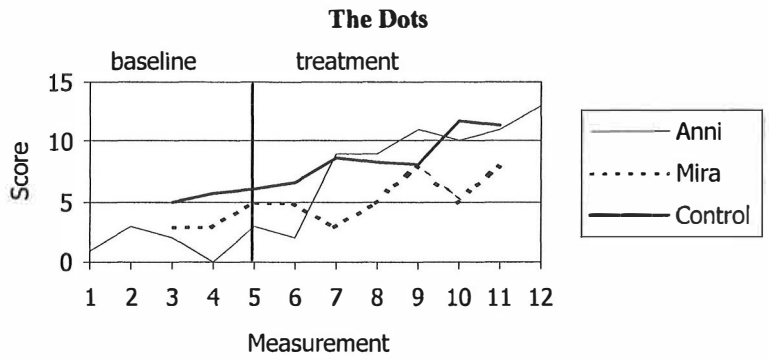
FIGURE 10 The results of the time-series measurement.The tests: the Tower of Hanoi, the Porteus Mazes test, the VMI test, the Underlining Test, the Dots, the Orientation and the Comparisons. The scale for the UL test presented on the right side and the scale for the other tests on the left side.

The pre-post measurements of WISC-R revealed that Mira improved her verbal

intelligence quotient (VIQ) from 68 to 74 and her performance intelligence quotient (PIQ) from 63 to 70. Anni did not show any improvement in the VIQ (from 96 to 97), but improved her performance IQ from 78 to 87. Mira showed improvement in most of the perceptual tests as well as in the tests measuring language and memory. A few subtests of the TVPS also showed some improvement. No improvement was found in the problem solving tests. Anni showed improvement in the problem solving tasks, motor performance, and two of the subtests of the TVPS. The perceptual, language and memory tests did not show improvement, see Figure 9. Thus, Mira showed improvement in most of the pre-post assessments which were used for assessing far-transfer effects, while Anni showed more modest gains which were limited to part of the functions trained in the intervention.

Detailed observation of the time-series data revealed that Mira improved her performance remarkably in only one of the no-transfer tasks (the Orientation). In the other no-transfer task (the Dots) there occurred steady but modest improvement throughout the assessments. No improvement was seen in the near-transfer tasks.
Anni showed steady improvement in one of the no-transfer tasks (the Orientation) and in two of the near-transfer tasks (the PM and the UL), and remarkable improvement in the other no-transfer task (the Dots) and in two of the near-transfer tasks (the TOH, the VMI), see Figure 10.

When the results of the time-series measurements of the participants were compared with the results of the control participants, it was revealed that Mira had improved her performance in the no-transfer tasks (the Dots, the Orientation) at about the same rate as the control participants, see Figure 11. Her performance in the near-transfer tasks was lower than that of the control participants already from the beginning, and she improved her performance slower than the controls, see Figure 12. Anni improved her performance more than the control group in the VMI test and in the two tasks requiring no-transfer (the Dots, the Orientation), but the improvement in other three tests (the TOH, the PMT and the UL test) was at the same level with the controls. When the performance of the participants was compared with the control group in a task, which was not part of the intervention (the Comparisons), the participants performed at the same level with the control group.



(Figure 11 continues on the next page)

(Figure 11 continues)

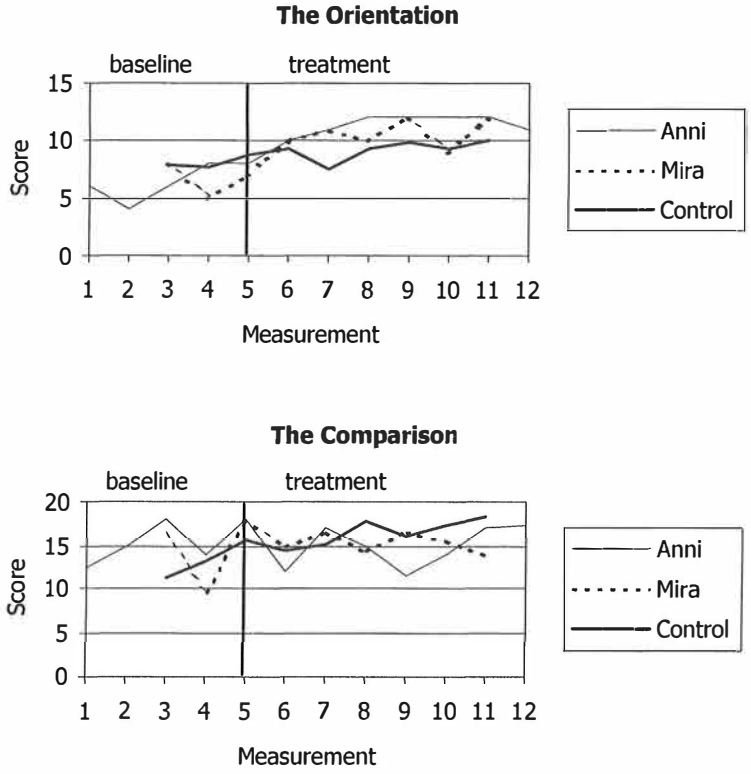
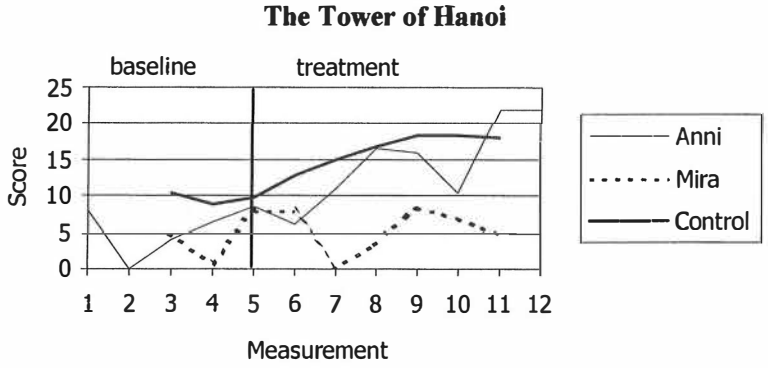
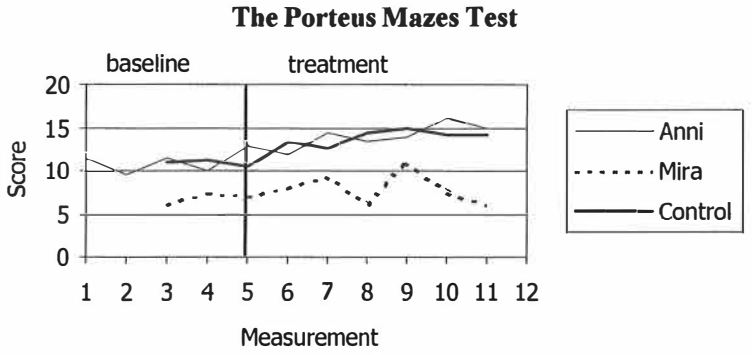
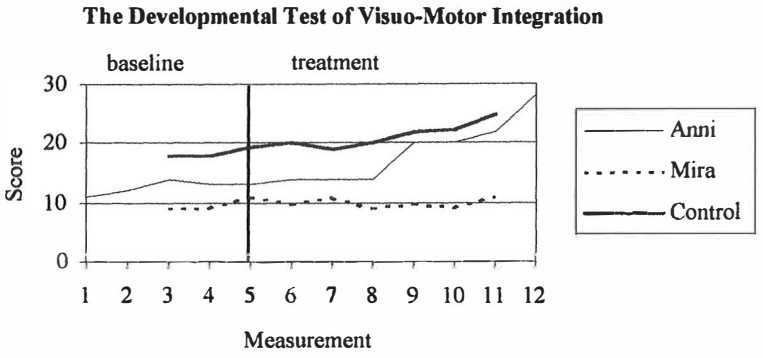


FIGURE 11 Comparison of the time-series data of the participants receiving intervention with the control group without intervention. No-transfer tasks: The Dots, the Orientation. Task unrelated to the intervention: the Comparisons.

The behavior questionnaires filled out by the teacher and parents were used for measuring the far-transfer of practice effects to daily activities. In the questionnaires, see Figure 13, Mira showed increase of desirable behavior and decrease of negative behavior. The teacher ratings showed almost no improvement in the behavior of Anni, and the amount of undesirable behavior even increased in some scales. The behavior evaluations of the parents showed some improvement.







(The Figure 12 continues on the next page)

(Figure 12 continues)

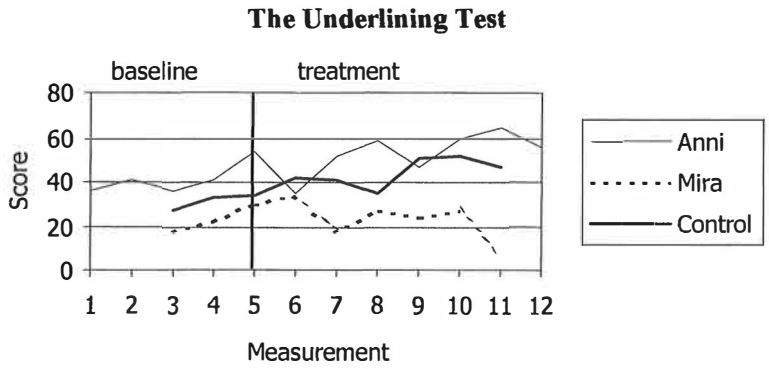


FIGURE 12 Comparison of the time-series data of the participants receiving intervention with the control group without intervention. The tests (near-transfer tasks): The TOH test, the PM test, the VMI test and the UL test.

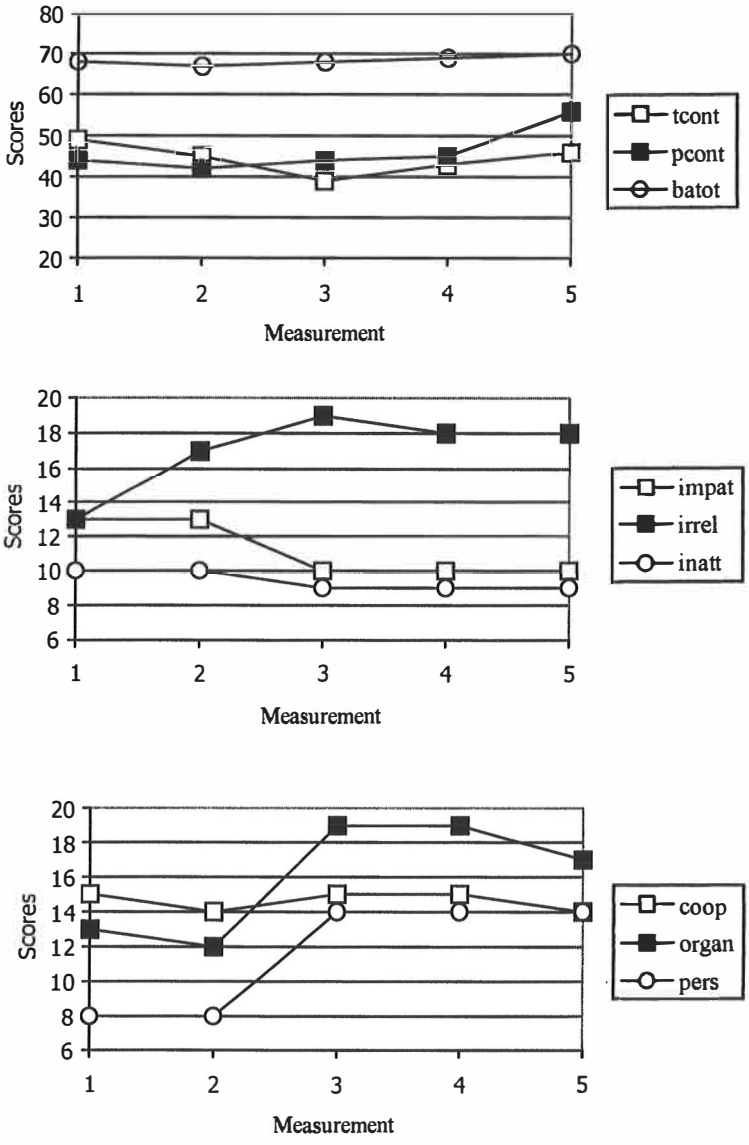
6.5 Discussion

The results of this experiment show that collecting and analyzing only one kind of data in single-case research can lead to biased conclusions. Mira improved her performance in most of the far-transfer tasks, including the WISC-R, in most of the neuropsychological tests, and in the behavior evaluations of the teacher and the parents. Additionally, her time-series data of no-transfer tasks showed some, although relatively modest improvement, but the near-transfer tasks failed to show any improvement.

The results of Anni seemed to be opposite. She showed almost no improvement in the neuropsychological tests requiring far-transfer effects of the intervention. Her time-series data, however, showed improvement both in the no-transfer and in the near-transfer tasks. The behavior questionnaires revealed an interesting difference between the assessments of parents and teacher. Parents found some improvement in her behavior, but the teacher reported either no meaningful change or increase in disruptive behavior during the intervention.

How should these various, partly contradicting findings be interpreted? This experiment does not provide enough evidence to prove that the improvement of Mira's performance in the pre-post assessment resulted from intervention effects. There are two reasons for this conclusion. First, there was only slight improvement of performance in the time-series data, and even that was restricted to the no-transfer tasks. Secondly, improvement in the neuropsychological pre-post assessment occurred evenly in almost all the tests. If intervention produces positive far-transfer effects without near-transfer effects, the reason for improved performance could be a halo effect, or regression to mean (Speer, 1992) because her pre-intervention results were low. The positive changes reported by the teacher and the parents, support a

Anni



(Figure 13 continues on the next page)

(Figure 13 continues)

Mira

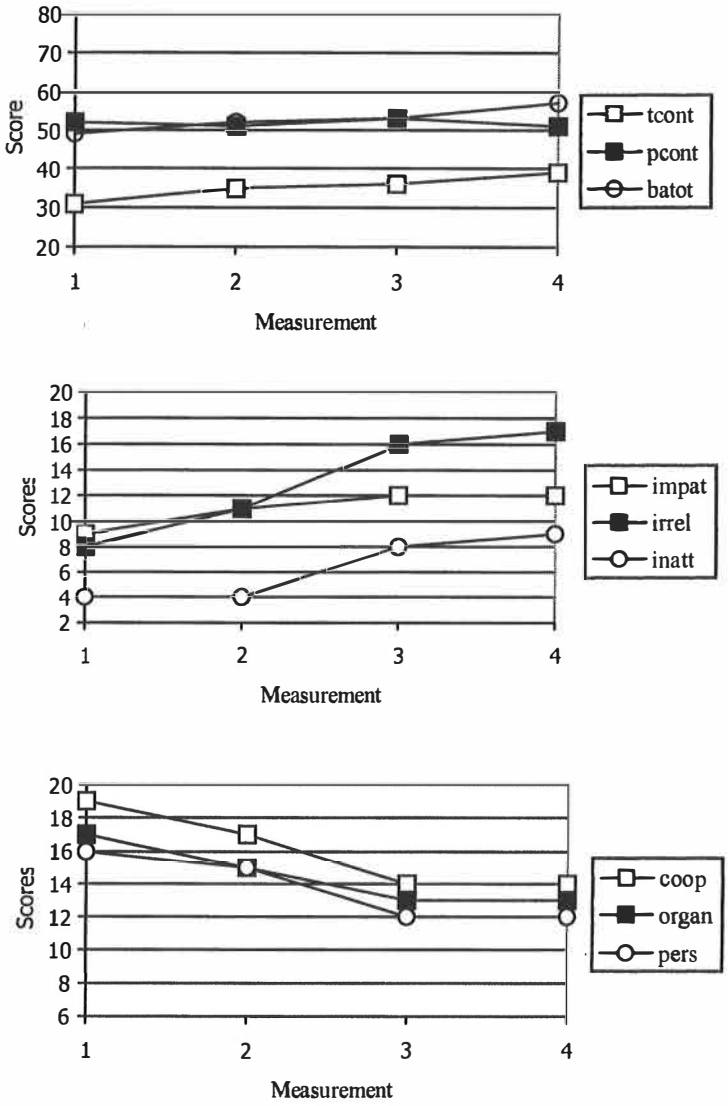


FIGURE 13 Results of the behavior questionnaires.The scales: the Teacher's Self-Control Rating Scale (filled by teacher), the Teacher's Self-Control Rating Scale (parents average), the Academic Performance Rating Scale (teacher), the subscales of co-operation, work organization, perserevance, impatience, irrelevant thinking, inattention from the Deveraux Elementary School Behavior Rating Scale (filled by teacher). In the first six scales, maximum of the score indicates desirable behavior, while in the last three scales, maximum of the score indicates undesirable behavior.

conclusion that the reason for improved performance could be improved motivation, positive attitude towards the clinician or tasks, or other unspecific effect of intervention or a genuine improvement of performance, which was completely unrelated to the intervention. Anni's opposite results showed improvement in the time-series measurement, but no improvement in the pre-post measurement and in the behavior evaluation of the teacher which could be interpreted as evidence of near-transfer effects, and lack of far-transfer effects of intervention.

Procedure-specificity of the treatment (Seron, 1997) was assessed by comparing the time-series data of no-transfer tasks with an intervention unrelated task (the Comparison). Improvement occurred also in the Comparison test, but the participants did not improve their performance more than the control group. Thus, this result could be an indication of a procedure-specific effect of the intervention at least regarding the no-transfer tasks.

The traditional analysis of the time-series data does not take practice effects into account. When neuropsychological or achievement tests are used with short test-retest intervals, practice effects could account for a great amount of improvement even if alternative versions of the tasks are used (Ahonniska et al.,c, in press.). Thus, in order to interpret the results of the time-series reliably, the time-series data of the participants was compared with the time-series data of a control group, which did not receive intervention. In this comparison, the intervention effects seemed remarkably smaller than in the traditional analysis. In Mira's case, only the no-transfer tasks showed improvement at the same level as the control groups. In all the other tasks, she benefited from the repeated testing less than the control group and the gap between her performance and the performance of the control group widened during the intervention. In Anni's case, the traditional analysis of the time-series data showed improved performance in six of the tests. When the time-series data were compared with the control data, only two of the no-transfer tasks and one of the near-transfer tasks (the VMI) showed greater improvement than the control group.

It can be argued that the results of the time-series of neurologically impaired participants should not be compared to the growth curve of the normal participants. Generally, neurologically impaired subjects benefit from repeated exposure to the same test less than the normal participants (Rapport, Brooke-Brines, Axelrod, & Theisen, 1997; Shatz, 1981). Thus, the amount of change which with the normal subjects would reflect practice effects could be a significant treatment effect with participants having learning disabilities. The time-series data of Mira might be evidence of this phenomenon. Her performance in the beginning of the assessment was low, and after the intervention only no-transfer tasks showed improvement at the same rate as the control group. In the near-transfer tests, her performance stayed on a lower level than that of the control group, and the gap between her and the control group even widened, in spite of the intervention. Thus, it could be assumed that the improvement of performance in no-transfer tasks which follows the practice effects of control group indicates significant intervention effects, and the performance in other tests indicates lack of near-transfer effects of intervention.

However, very little is known about practice effects in time-series measurement, and almost nothing about the practice effects among neurologically impaired participants. Although the neurologically impaired participants possibly show fewer practice effects than the normal participants, this assumption might easily lead to underestimating the influence of practice effects and to being overly

optimistic in regard to intervention effects. Thus, if psychological tests are used as an assessment tool in time-series measurement, separating intervention effects from practice effects requires growth curves of various control and patient groups (see also Denckla, 1994).

As a conclusion, time-series analysis seems to be the most reliable method for assessing change. However, time-series data might also provide overly optimistic results because of practice effects. Thus, if psychological or achievement tests are used in time-series analysis with short assessment intervals, it is recommended to create growth curves of the same tests without intervention.

7 GENERAL DISCUSSION

7.1 Separating practice effects from development and intervention effects

7.1.1 The effect of age on practice effects

When change in children's cognitive functions is assessed by time-series measurement, interpretation of data becomes complicated. Improved performance may result from variations in the dependent variable, e.g. intervention method, but also from development or practice effects. Separating the intervention effect from all the other variables requires estimating the effect of development and practice correctly.

The previous studies examining practice effects have repeated the assessment only once or twice (see e.g., Casey, Ferguson, Kimura, & Hachinski, 1989; Dyche & Johnson, 1991a; McCaffrey, Ortega, & Haase, 1993; McCaffrey et al., 1992a; McCaffrey et al., 1992b; McCaffrey et al., 1995; Neyens & Aldenkamp, 1996; Rawlings & Crewe, 1992; Tuma & Appelbaum, 1980). This research is the first one to study practice effects of neuropsychological tests repeating the assessments several times with relatively short test-retest intervals. It provided growth curves in two age groups for two tests of executive functions, one test of visuo-motor integration, and one test of visual discrimination. The magnitude of practice effect in various tests was analyzed, as well as the effect of age on the practice effect. The results indicated that the visuo-motor test did not show practice effects while the other tests did. Age had an effect on the results. In most of the tests, the older children showed larger practice effects than the younger children.

In addition to the larger practice effects in the older than the younger group, the older participants improved their performance remarkably faster than the younger group in all other tests except for the VMI test. Thus, the more developed cognitive abilities of older children had apparently improved their learning ability and helped them to use repeated assessments as learning opportunities more efficiently than the younger children. A rather similar phenomenon has been called "the rich get richer", as the participants with higher than average level of intelligence show larger practice effects in intelligence tests than their lower than average controls (Rapport, Brooke Brines, Axelrod, & Theisen, 1997). There are also reports showing that elderly adults and brain injured patients show no practice effect or significantly smaller practice effects than

normal controls (Mitrushina & Satz, 1991; Satz, 1981). These results emphasize the importance of adequate selection of the control group providing growth curves. The issue will be discussed in detail later.

7.1.2 The effect of alternative versions and domain of measurement on practice effects

Alternative versions of tests are supposed to prevent or diminish practice effects (Denckla, 1996). In fact, alternative versions have been reported to prevent practice effects in learning of word lists (Crossen & Wiens, 1994; Parker, Eaton, Whipple, Heseltine, & Bridge, 1995). No previous reports exist about the effect of alternative versions on the practice effects in tasks measuring executive functions or visuo-perceptual functions. In this experiment, alternative forms of the same difficulty level were used in one of the executive function tests and in a visual discrimination task. In either of the tests, alternative forms did not prevent practice effects, although the practice effects might have been even larger without the alternative versions. However, the visuo-motor task, which did not have alternative forms, did not show a practice effect. Thus, it can be concluded that the domain of cognitive processing required in the task, and the method of measuring that particular cognitive domain determine the magnitude of practice effect more than the variation in the specific contents of the test.

Familiarity with the cognitive operations required by the test also might explain, why certain tests are more sensitive to practice effects. Probably, if the cognitive operations required by the test are seldom practiced in normal life, repeated assessments are used as learning opportunity in the first few assessments. During the assessments the participants learn to generate task-relevant strategies which makes application of those strategies to slightly different content relatively easy. This results in improvement of performance in most tests, and especially in executive function tests. In the cases of fast improving growth curves, the result of first assessment might only indicate the ability of the participant to perform in a unfamiliar situation, with weakly understood requirements. The real ability of the participant could rather be seen after the growth curve slows down to even level. If the cognitive operations required by the test are practiced regularly in the daily life, it results to overlearning of the operations, and lack of practice effects in test situations.

7.1.3 How to control the amount of practice effects

When neuropsychological tests are used for assessing change in repeated measurement, interpretation of the results would be the least problematic if the magnitude of practice effects could be minimized or controlled. There are several possibilities to minimize practice effects. First, cognitive functions could be assessed by tests which are known to be insensitive to practice effects. These include, for example, various tests of visuo-motor coordination, motor steadiness, reaction time, and selective attention (McCaffrey, Ortega, & Haase, 1993). Second, one could also to use tests in which alternative versions are known to prevent practice effects, i.e., word list learning (Crossen & Wiens, 1994; Parker, Eaton, Whipple, Heseltine, & Bridge, 1995). Probably, in the tests measuring cognitive abilities like reading and mathematics there are more possibilities to created

alternative versions which would be less sensitive to practice effects than the neuropsychological tests. However, all these aforementioned tests represent only a few cognitive domains. Most of the commonly used tests show significant practice effects, or the magnitude of practice effects is not known at all. More research is needed for assessing the sensitivity of commonly used neuropsychological tests to practice effects.

Third, if the theoretically interesting cognitive domain does not have tests which would be insensitive to practice effects, detailed analysis of performance in commonly used tests might enable creating new and more reliable scores. A few existing examples of these kinds of scores could be the perserative errors score (Denckla, 1996, referring to personal communication of Pennington) or percent of conceptual level responses (Heaton & al., 1993; Tate, Perdices, & Maggiotto, 1998) in the Wisconsin Card Sorting Test, and the relative time scores or new error scores in the Tower of Hanoi Test (Study II). The validity and reliability of these aforementioned scores has to be assessed further. However, they provide an interesting example for analyzing in detail various strategies which participants use for solving or attempting to solve commonly used neuropsychological tasks. At its best, this method could shift the focus of research from assessing the result towards analysis of process of performance. In fact, clinicians commonly analyze the qualitative aspects of test performance, and process analysis is only a quantification of this analysis (see e.g., Kaplan, 1988).

A fourth method for minimizing practice effect is to create increasingly difficult versions of the same test, following the example of the Porteus Mazes Test. In this experiment, the Porteus Mazes succeeded in preventing practice effects with its three versions, but in time-series analysis, a few more alternative versions could be needed. Difficulty level of many kinds of neuropsychological tests can be varied by manipulating the amount of stimuli and the time load of the task. Recent research of executive functions has suggested manipulating working memory load and demands of prepotent response inhibition in order to create variations in processing demands of the tasks (Roberts & Pennington, 1996). Increasing working memory load could be achieved by increasing of the length of time one must keep information on line and by introducing an interfering task (Roberts & Pennington, 1996). Creating alternative versions of increasing difficulty requires a lot of research and preliminary assessment with several age groups. It possibly requires as much work as developing a totally new test.

Performance in executive function tasks is affected strongly by familiarity with the problem and the test structure (Denckla, 1994; Roberts & Pennington, 1996), and domain specific knowledge and experience (Torgesen, 1994). Even if the executive functions were defined relatively narrowly as mechanisms of working memory and inhibition, "understanding the problem and knowing the correct answer would contribute to inhibiting or ignoring misleading tendencies" (Roberts & Pennington, 1996, p. 119). Thus, the fifth possibility of minimizing practice effects might include tasks in which the assessed cognitive ability is the same but both the structure and content of the test is different (Denckla, 1996). This method would require even larger modifications to the existing tests than the previously discussed alternatives. Whether and how this interesting theoretical suggestion could be operationalized is unknown.

In conclusion, it is very difficult to prevent practice effects altogether or to diminish them significantly in most of the commonly used tests. Executive function

tests which measure the use of cues or response to feedback have been suspected of showing large practice effects (Denckla, 1994), but this experiment shows that simple tests of visual discrimination also show significant practice effects even if alternative versions are used. In order to use these tests in repeated assessment, one has to be able to estimate the magnitude of practice effects. This could be done by providing growth curves of the relevant tests with an appropriate amount of assessments and test-retest intervals using reference groups which would be similar enough to the target population in the critical features (age, cognitive abilities, etc). The growth curves could help to determine the type and amount of change in performance which could be accepted as therapeutic gain (Denckla, 1994).

7.1.4 Separating practice effects from intervention effects

This experiment provided growth curves for four tests in two age groups. A lot more research has to be done in order to provide enough growth curves for theoretically interesting tests and relevant age groups. The test manual providing information on practice effects of widely used assessment instruments is an extremely valuable starting point in this direction (McCaffrey, Duff, & Westervelt, 2000), although most of the practice effect data provided in the manual is repeated only once or twice, which is far less frequently than needed for reliable interpretation of time-series analysis.

The intervention study with single-case paradigm (Study IV) presented an example of a case-study which assessed effects of an intervention method targeted to improve visuo-perceptual functioning in two children. The assessment of intervention effects in single-case research included pre-post measurement with various psychological or neuropsychological tests, behavior questionnaires, and time-series analysis with tasks which were related to the contents of the intervention with various degrees (no-transfer, near-transfer, far-transfer). The experiment showed that various assessment methods might easily yield to contradictory results for various reasons. The results of pre-post measurements are easily affected by variations in such factors as co-operation, motivation and attention. The behavior questionnaires could be affected by emotional and interpersonal reasons completely unrelated to intervention method. The time-series is probably the most reliable method because random errors can be seen as deviations from the general trend of the data. However, in order to separate practice effects from intervention effects, the results should be compared with growth curves of the same tests measured in a suitable control group. The results of this experiment show that if the practice effects are not thoroughly analyzed, the intervention effects may easily be overestimated and practice effects underestimated. As long as the practice effects, and development are not separated from intervention effects, intervention research will continue publishing too optimistic research reports in various fields.

7.2 Future implications for intervention research

7.2.1 Problems in demonstrating procedure-specific effects with neuropsychological tests

Assessing change in single-case research with neuropsychological tests has several advantages. Neuropsychological tests are theoretically based, they provide normative data for various age groups and information about the validity and reliability of the test. Occasionally, some information exists even about the practice effects found in repeated assessment. Neuropsychological tests are also familiar to other professionals and their use needs not to be justified.

However, assessing change in time-series research with neuropsychological tests has serious disadvantages. Clinical reasons such as allocation of time often limit the amount of repeated assessments. Repeating neuropsychological assessment is severely restricted by the lack of alternative forms and by practice effects even if alternative forms are used. Thus, neuropsychological assessment very rarely can be repeated frequently enough to provide a reliable trend for visual inspection or to allow using statistical methods. Both of these reasons diminish the reliability of interpretations (see e.g., Robey, et al., 1999).

Additionally, neuropsychological measures are commonly not very sensitive measures for assessing change. In order to attribute improvement of the performance unequivocally to certain treatment, the treatment effects should be procedure-specific. This means that major improvement should occur in the tasks directly related to intervention (no-transfer), minor improvement should occur in tasks slightly differing from the intervention tasks (near-transfer), and no improvement should be seen in tasks unrelated to the intervention (Seron, 1997). If neuropsychological tests are used in assessing the intervention effects, the strategies and abilities, which were trained in the intervention, have to generalize to abstract material requiring far-transfer in order to show positive effects of the intervention.

According to a fundamental assumption of cognitive neuropsychology, the cognitive systems are organized in a modular way (Fodor, 1983), i.e. cognitive processes are composed of subparts which are discrete and relatively independent (Temple, 1998). Fodor's original idea of modularity argues that executive functions do not fractionate but are common across cognitive systems (Fodor, 1983). Although further research has opposed this restrictive application of modularity concept arguing that executive systems are also potentially fractionable (see e.g. Temple, Carney, & Mullarkey, 1996), most of the executive function tests might be less pure measures, showing more correlation with other test results than for example, language and visuo-spatial processes (Burgess, 1997). This means that strictly limited, procedure-specific effects of interventions might be more difficult to prove in the case of executive functions than in language.

If intervention, which aims to improve executive functions, shows improved performance as far from intervention material as in neuropsychological tests measuring executive functions, improved performance probably can be seen in a number of other tests as well because improved and generalized problem solving strategies probably improve performance on several areas of cognitive functioning. Thus, proving procedure-specific effects of any intervention methods, including

rehabilitation of executive functions, requires creation of intervention related tasks in addition to neuropsychological tests. For example, the target of the intervention could be to learn to control one's behavior in problem solving situations by verbalizing essential components of performance (Meichenbaum & Goodman, 1971), and to learn to apply it in puzzle construction. Thus, the intervention related tasks requiring no transfer should measure the ability to control one's behavior in the same or very similar puzzles as used in the intervention.

7.2.2 Influence of assessment methods on intervention methods

In intervention research, complicated interaction exists between the methods of measuring change (assessment) and the methods of creating change (intervention). Intervention methods should be based on neuropsycological theory about the nature of disorder, the intact abilities, the theoretical understanding about the means of change, and the needs of the patient and his/her immediate environment (Kazdin, 1997, Basso & Marangolo, 2000). Assessment methods need to be chosen according to the goals of the intervention, availability, reliability, and sensitivity of the tests. However, setting the goals of the intervention has commonly been influenced more by manifestation of the disorder in the neuropsychological assessment than by practical individual needs or theoretical understanding of the phenomenon. For instance, if the child has visuo-perceptual problems, which are manifested as deficits in low performance in the Developmental Test of Visuo-Motor Intergration (Beery (1989) and the Visual Discrimination and the Figure-ground Discrimination subtests of the Test of Visuo-Perceptual Skills (Gardner, 1982), the intervention is commonly targeted improving his/her performance in various copying tasks and visual discrimination tasks, and little time is allocated in training him/her to generalize the trained skills to visuo-perceptual problems which he/she meets everyday with school material.

The dominant influence of the assessment methods on the contents and goals of intervention methods is also manifested in several intervention programs (e.g. Frostig, Instrumental Enrichment) which are surprisingly similar to the traditional assessment methods. Positive results of these interventions measured by commonly used tests could be considered near-transfer effects, meaning effects requiring little or no generalization. There are numerous reports of clinical interventions (see e.g., Ahonniska & Aro, 1999; Aro, Mäntynen & Poikkeus, 1998; Davis & Coltheart, 1999; Evans, Emslie & Wilson, 1998) detailed programs of intervention methods (German, 1993; Hänninen, Kaukovalta, & Kuikka, 1985) and also some computer programs (e.g., Ahonniska, Strömmer, & Viholainen, 1998) which are based on the neuropsychological theory of the dysfunction instead of close resemblance with psychological assessment methods. However, efficacy of most of these theoretically based interventions needs to be thoroughly evaluated.

The influence of assessment methods on intervention methods has been relatively dominant leading to underestimating the importance of practical and academic manifestations of neuropsychological deficits. Because the efficacy studies of various intervention methods for perceptual-motor functions have shown only modest results (see e.g., Kavale, 1990a; 1990b; Kavale & Dobbins, 1993; Robertson et al., 1990; Robertson, 1997) and the modest improvement generally has been limited to intervention related tasks requiring no transfer (Robertson, 1997), it has been suggested that the target of the interventions should be improving the precise

academic skills (Hinshaw, 1992) or practicing cognitive functions closer to concrete problematic situations (Robertson, 1997). As an example, Robertson suggests that neglect patients should learn scanning during the course of reading or pushing the wheelchair, and not by sitting in front of a computer or by doing abstract puzzles. He also suggests: "Activities such as shape and color matching, cube copying, picture arrangement and sequencing, and similar abstract tasks should therefore be abandoned until their utility can be demonstrated" (p. 181).

What could it mean in practice if the intervention methods were shifted closer to concrete problem solving situations in daily activities? Specifically, in children with perceptual problems, which are often accompanied by problems in executive functions, the shift could mean learning the needed strategies with easy and interesting material, and practicing them in all difficult areas which are relevant for school learning and everyday life. Perception of directions, for example, could be practiced first with directions of silhouettes of animals and faces, then with directions of letters and numbers, then with the direction of reading and calculating processes, in drawing, and in orientation from one place to another in a two-dimensional environment (e.g., computer games) and in map reading and orienting in a real environment. In addition to special tasks designed for each participant, the intervention material should also contain exercises in the problematic school subjects using the normal textbooks or educational computer programs of the child. This shift would mean very close co-operation with parents and teachers in order to know, what kind of practical problems the particular cognitive deficit may create and to develop tasks that are synchronized in content and time with the classroom teaching.

If the intervention material and intervention context were moved closer to everyday activities, this would mean changes in assessment methods. Assessing intervention effects could include, not only neuropsychological tests, but also tasks which are closely related to intervention material, goal-oriented analysis (Kiresuk, Smith, & Cardillo, 1994), and behavior observation. This could mean laborious creation of tasks related to intervention with various degrees because sensitive, valid and reliable tests are not easy to design. However, the intervention-related tasks and goal-oriented tasks created specially for measuring effects of certain intervention probably would measure the procedure-specific effects of intervention better than neuropsychological tests. Important activities (e.g., tying shoelaces or making a puzzle) or child's performance of intervention tasks could be recorded on video with certain intervals and used later for behavior observation. Behavior observation has been used surprisingly little in neuropsychological assessment, although in the best case, video observation could be used for observing changes in behavior in intervention sessions with little disturbance. Definitely, neuropsychological tests are useful in time-series measurement as well as in pre-post measurement, but they should not be the only methods for assessing intervention effects.

7.3 Methodological considerations

In this study, the reliability values of the tests were assessed by the LISREL analysis instead of the commonly used test-retest reliability measured by Pearson correlation coefficients. The LISREL analysis is less used in reliability analysis, which makes it

more difficult to interpret and compare the results with other studies. However, the LISREL method has some important advantages in frequently repeated assessment compared to the traditional test-retest reliability. The test-retest reliability measured by correlation coefficients considers the differences between the subjects in the measured ability to be completely consistent from one measurement to another. This means that each child would be able to use the repeated assessment as a learning opportunity in a similar way and each child would improve his scores in a rate fitting to his/her first assessment. In reality, however, the rate of learning differed among children independent of the result of the first assessment. E.g. in the cluster analysis of the TOH test three groups of children could be differentiated. One which improved the performance very fast in the first few assessments, second, which improved very slowly and third which showed wide variation in the performance throughout the assessments.

This variation in the learning is taken into account in the LISREL analysis using simplex model, while it allows variation in the differences between the subjects from one measurement to another (stability can be lower than 1.0). It describes how well the test measures certain ability independent of the intra-individual consistency of the results from one measurement to another. The stability value of the LISREL analysis indicates how constant the differences are in the measured ability between subjects.

Reliability values of the tests of visuo-motor integration and visual discrimination were good in all the assessments. The reliability values of the problem solving tests were lower, although the reliability of the Tower of Hanoi Test improved after the first two assessments. The low reliability of the Porteus Mazes Test was probably due to the ceiling effects. The low reliability values of the Tower of Hanoi might be improved by changing the methods of presenting and scoring of the test, as discussed in the studies I and II. However, Denckla (1994) pessimistically suggests that "developmental sensitivity is more achievable condition than test-retest reliability with respect to executive function" (1994, p. 122).

In the study IV, choosing control group for the two participants on the intervention study was done according to the similarity of abilities in some tests, not according to the age. The control group was consisted of normal children, whose learning ability probably was better than that of the two subjects. Thus, it is possible that the practice effects in the control group might have been larger than the practice effects shown by the two participants of the intervention study. Although it is very difficult to form enough homogeneous groups of patient populations for group studies, in the future research the magnitude of practice effects should be assessed also in some groups of children patients, which vary in age and domain area of cognitive deficits. Also, the study II, suggesting revised methods for the Tower of Hanoi, had participants with various cognitive deficits, and the relevance of the suggested scoring method needs to be evaluated in normal participants.

This study presents growth curves of four tests. Two of them were tests measuring executive functions, one measuring visual discrimination and one visuo-motor functions. Although this study concluded that the visuo-motor test did not show practice effects and the other tests did, the results could have been different if executive functions and visuo-motor functions had been measured with other tests of the same cognitive domain. Thus the practice effects might have been found if visuo-motor functioning had been measured with the Rey Complex Figure Test

instead of the VMI test. Further research is needed in order to analyze, which variables of various tests increase or decrease practice effects in repeated assessment.

The rehabilitation method, the Instrumental Enrichment, is an example of abstract material, far away from daily activities in classroom and closer to methods of assessment of cognitive functioning. The advantages of this method are the emphasis of using accurate verbal labeling in solving complicated perceptual tasks, creating complicated problem solving strategies, and the aim to generalize learned principles to daily activities. The generalization to everyday activities was performed in discussions, where applications of the learned rules and strategies were searched from the scope of everyday life. Sometimes, examples for generalization of learned problem solving strategies were based on the concrete tasks of various schoolbooks.

However, the method is not based on explicitly and unequivocally defined theory, but rather, on heterogeneous concepts, whose empirical conformability is questionable (Buchel & Scarnhorst, 1993). The complicated and partly out dated terminology makes learning and applying the intervention method quite laborious. Because the intervention material is not based on recent theory of neuropsychology and cognitive psychology, it is not easily applied to known diagnostic categories and the suggested areas of influence are wide and conceptually ill-defined. The efficacy of the method or the relevancy of the content of the exercises has not been proven unequivocally. Wide generalization of the strategies would have needed more systematic use of general textbooks, in addition to discussion and examples derived form the school subjects, closer co-operation with the teachers or intervention performed in the classroom by the teacher.

7.4 Concluding remarks

The current results demonstrate that although in single-case research the time-series assessment is the most reliable method for assessing intervention effects, even the results of time-series assessment might seriously be limited by developmental and practice effects. Various tests show different magnitude of practice effects, even if the alternative versions are used. Age also has an influence on practice effects. In most of the commonly used tests, minimizing practice effects by creating alternative scoring, alternative versions, or alternative formats of the tests is laborious and requires a great amount of research. Most likely, the easiest method separating practice effects from intervention and development effects is to provide growth curves for relevant tests. There is a great need for analysis of practice effects in various age groups and in different patient groups.

Although neuropsychological tests have their indisputable advantages in normal neuropsychological assessment, using them in repeated assessment is problematic. Neuropsychological tests lack alternative forms which limits their usefulness in repeated assessment due to practice effects or limitations in test validity. Additionally, neuropsychological tests cannot measure procedure-specific effects of intervention, but rather more generalized effects in relatively far-transfer tasks. In order to prove intervention to be procedure-specific, we need to generate intervention related tasks, goal-oriented tasks, and a wide application of behavior

observation methods.

Intervention research and assessing change in general needs both single-case and group research because both of the methods have their assets and limitations. Analyzing change presents challenges, which require analysis of practice effects and test-retest reliability, and require new ways of measuring and analyzing performance. Developing methods of repeated assessment mainly improves time-series analysis in single-case research, but revised scoring and growth curves and could be used in group research as well.

The relationship between the therapist and the subject is always inseparable from the intervention; it varies in every individual case and affects the cognitive results of intervention. The improvement or lack of improvement might result from reasons that are not cognitive, but motivational, affective, or cultural by nature (Seron, 1997). However, both the patients and the clinicians need intervention methods, which also rely on neuropsychological and cognitive theories and not only on personal charisma of the therapist. Thus, the most important goal of intervention research is to gain knowledge of various factors influencing change, to separate efficient methods from futile ones, and to create more efficient intervention methods. Good intervention research also enhances neuropsycholgical theory formation because proper understanding about the mechanisms of rehabilitation of functions would help to understand normal cognitive functioning and their disorders.

YHTEENVETO

Muutoksen mittaaminen on haasteellinen psykologisen arvioinnin osa-alue, joka edellyttää mittausten toistamista. Kliinisessä neuropsykologiassa toistomittausta käytetään esim. seurattaessa riskilasten kehitystä, etenevien sairauksien vaikutuksia, spontaania paranemista trauman jälkeen ja tutkittaessa kuntoutuksen vaikutuksia. Pääasiallisesti toistomittauksen menetelmiä käytetään yksittäistapaustutkimusasetelmissa, mutta mittauksia toistetaan myös ryhmätutkimuksissa. Sekä yksittäisettä ryhmätutkimusasetelmilla on etunsa ja haittansa muutoksen arvioinnissa. Tämä tutkimus keskittyy toistomittausongelmien analysoimiseen yksittäistapaustutkimuksessa ja harjoitteluvaikutuksen erottamiseen kehityksen ja intervention vaikutuksista.

Yksittäistapaustutkimuksen toistomittausasetelma eli aikasarja-analyysi soveltuu hyvin muutoksen mittaamiseen kliinisen työn yhteydessä, koska aikasarja-analyysi ei edellytä satunnaista koehenkilöiden valintaa eikä koeryhmien keskinäistä yhdenmukaisuutta. Aikasarja-analyysi soveltuu kliiniseen käytäntöön myös siitä syystä, että se vaatii pienempiä muutoksia tavanomaisiin kuntoutus- ja mittausrutiineihin kuin ryhmätutkimus. Aikasarja-analyysi antaa myös yksityiskohtaista tietoa muutoksen ajoittumisesta ja sen syistä. Aikasarja-analyysillä on kuitenkin rajoituksensa. Koska aikasarja-analyysi voidaan tehdä mitaten vain yhden tai muutaman koehenkilön suorituksia, luotettavien ja yleistettävien päätelmien tekeminen edellyttää usein toistettuja mittauksia samoilla koehenkilöillä. Tämän seurauksena tulokset saattavat olla tavanomaisesta poikkeavia ja niiden yleistettävyys on mahdollisesti rajoitetumpi kuin ryhmätutkimuksen tulosten yleistettävyys. Aikasarja-analyysissä tulosten tilastollista merkitsevyyttä on vaikea arvioida. Mittauksen toistaminen vaikuttaa myös helposti testin validiteettiin ja toistomittausreliabilitettiin.

Aikasarjamittaus on alunperin kehitetty mittaamaan käyttäytymisen muutoksia ja menetelmä onkin soveltuvin käyttäytymisen havainnointiin. Aikasarja-analyysi sopii myös sellaisille alueille, joissa yksilön käyttäytyminen on sangen pysyvää. Tällaisia ovat esim. psykofysiologiset tutkimukset. Aikasarja-analyysia käytetään kuitenkin yhä enemmän myös sellaisessa psykologisessa ja neuropsykologisessa tutkimuksessa, joissa riippuvana muuttujana ei ole tietyn käyttäytymisen yleisyys, vaan psykologisen tai neuropsykologisen testin tulos. Testitulokset eivät ole ainoastaan herkkiä muutoksen mittareita, vaan ne ovat alttiita myös harjoitteluvaikutukselle. Testien harjoitteluvaikutuksia ei juurikaan tunneta, mutta niiden määrään vaikuttavat monet, suurimmalta osalta tuntemattomat tekijät, kuten koehenkilöiden ikä, sukupuoli ja kognitiivinen toimintataso, testin sisältö ja mittauskohde, vaihtoehtoisten versioiden saatavuus, mittausten määrä ja mittausvälin pituus.

Tässä tutkimuksessa tarkastellaan aikasarja-analyysin ongelmia kouluikäisten

koehenkilöiden muutosta arvioitaessa. Tutkimuksen koehenkilöjoukko muodostui kahdesta yleisopetuksessa olevasta oppilasryhmästä. Nuoremmat oppilaat olivat tutkimuksen alussa 7.7 vuotiaita ja vanhemmat oppilaat 11.6 vuotiaita. Tutkimuksessa käytettäviä testejä toistettiin yhdeksän kertaa puoletoistavuoden aikana niin, että mittausväli oli kaksi kuukautta. Aluksi tämä tutkimus tarkastelee yksityiskohtaisesti yhden toiminnanohjausta mittaavan testin (Hanoin torni) erilaisia pisteytysmahdollisuuksia, niiden reliabiliteettia ja stabiliteettia toistomittauksessa, sekä ehdottaa uusia tapoja Hanoin tornin pisteyttämiseksi. Tutkimus antaa myös kasvukäyrät yhdelle visuomotoriselle testille, yhdelle visualisen tarkkuuden testille ja kahdelle ongelmanratkaisutestille. Tutkimuksessa tarkastellaan myös, miten ikä, kognitiivinen taso ja vaihtoehtoisten versioiden käyttö vaikuttavat harjoituksvaikutuksen määrään. Tämän lisäksi tutkimus esittelee esimerkin kuntoutustutkimuksesta, jossa koehenkilöiden toiminnan muutosta tarkastellaan alku-loppumittausten, käyttäytymisarviointien ja aikasarja-analyysin avulla. Aikasarja- analyysien tulosta verrataan kontrolliryhmän tarjoamaan kasvukäyrään.

Tämä tutkimus on ensimmäinen, joka on tutkinut neuropsykologisten testien harjoitteluvaikutuksia toistamalla mittauksia useita kertoja suhteellisen lyhyellä mittausvälillä. Tutkimus tarjoaa kasvukäyrän useampaan neuropsykologiseen testiin. Tutkimuksen tulokset osoittivat, että visuo-motorinen tehtävä ei ollut herkkä harjoitteluvaikutukselle, kun taas ongelmanratkaisutehtävissä ja visuaalisen tarkkuuden tehtävässä ilmeni voimakasta harjoitteluvaikutusta. Rinnakkaiset versiot eivät poistaneet harjoitteteluvaikutusta. Vanhemmat koehenkilöt osoittivat suurempaa harjoitteluvaikutusta kuin nuoremmat koehenkilöt.

Tutkimuksen perusteella voidaankin päätellä, että luotettavien johtopäätösten tekeminen aikasarja-analyysissä edellyttää harjoitusvaikutusten minimointia tai ainakin niiden erottamista kehittymisestä ja kuntoutusvaikutuksista. Aikasarja-analyysin päätelmiä voidaan tehdä luotettavammaksi käyttämällä testejä, jotka eivät ole herkkiä harjoitusvaikutuksille, käyttämällä rinnakkaisia versioita testeissä, joissa nämä versiot estävät harjoitteluvaikutuksen, luomalla yksityiskohtaisen analyysin perusteella uusia, harjoitusefektille epäsensitiivisiä pisteytystapoja olemassa oleviin testeihin ja luomalla vaikeutuvia versioita samasta testistä. Aina harjoitusvaiktusten poistaminen tai vähentäminen ei ole mahdollista, ja tällöin harjoitusvaikutuksen määrää voitaisiin arvioida kasvukäyrien avulla, kuten tässä tutkimuksessa on tehty. Kaiken kaikkiaan, neuropsykologisten mittareiden käyttö toistomittauksessa voi olla ongelmallista, ja interventiotutkimuksessa saattaisikin olla hyödyllistä täydentää neuropsykologisia testejä mittausmenetelmillä, jotka ovat lähellä interventiomenetelmää, tai intervention päämääriä.

REFERENCES

Ahonniska, J. & Aro, T. Hahmotusvaikeuksien kuntoutus. In T. Ahonen, & T. Aro (Eds) *Oppimisvaikeudet. Kuntoutus ja opetus yksilöllisen kehityksen tukena* (pp. 102-119). Juva:Atena.

Ahonniska, J., Ahonen, T., Aro, T., & Lyytinen, H. (2000b). Suggestions for revised scoring of the Tower of Hanoi task. *Assessment, 7,* 311-320

Ahonniska, J., Ahonen, T., Aro, T., Tolvanen, A., & Lyytinen, H. (2000a). Repeated assessment of the Tower of Hanoi task: Reliability, and age effects. *Assessment, 7,* 297-310.

Ahonniska, J., Ahonen, T., Aro, T., Tolvanen, A., Lyytinen, H. (submitted) Practice effects of visuo-motor and problem solving tests in children.

Ahonniska, J., Strömmer, K., Viholainen H. (1998). *Vista. Käsikirja.* Helsinki: Opetushallitus.

Anderson, P., Anderson V., & Lajoie, G. (1996). The Tower of London Test: Validation and standardization for pediatric subpopulations. *The Clinical Neuropsychologist, 10,* 54 - 65.

Aro, T., Mäntynen, H. & Poikkeus, A-M. (1999) Dysfattisten lasten ryhmäkuntoutus toiminnanohjauksen ja itseilmaisun harjaannuttamiseksi. *NMI-Bulletin, 9,* 22-31.

Basso, A., Marangolo, P. (2000). Cognitive neuropsychological rehabilitation: The emperor's new clothes? *Neuropsychological Rehabilitation, 10,* 219-229.

Barlow, D., & Hersen, M. (1985). *Single-case experimental designs: Strategies for studying behavior change. 2nd ed.* New York: Pergamon Press.

Becker, M.G., Isaac, W., & Hynd, G.W. (1987). Neuropsychological development of nonverbal behaviors attributed to "frontal lobe" functioning. *Developmental Neuropsychology, 3,* 275-298.

Beery, K.E. (1982, 1989*). Revised administration, scoring and reaching manual for the developmental test of visual-motor integration.* Cleveland: Modern Curriculum Press.

Borys, S.V., Spitz, H.H., & Dorans, B.A. (1982). Tower of Hanoi performance of retarded young adults and nonretarded children as a function of solution length and goal state. *Journal of Experimental Child Psychology, 33,* 87-110

Buchel, F.B, & Scarnhorst, U. (1993). The Learning Potential Assessment Device (LPAD): Discussion of theoretical and methodological problems. In J.H. Hamers, K. Sijtsma et A.J.J.M. Ruijssenaars (Eds.) *Learning Potential Assessment. Theoretical, methodological and practical issues.* Amsterdam: Swets & Zeitlinger.

Burgess, P. (1998). Theory and methodology in executive function research. In P. Rabbit (Ed.*) Methodology of Frontal and Executive Function.* Hove: Psychology Press (pp. 81-116).

Bushke H., & Fuld, P.A. (1974). Evaluating storage, retention, and retrieval in disordered memory and learning. *Neurology, 11,* 1019-1025.

Caplan D. (1988). On the role of group studies in neuropsychological and pathopsychological research. *Cognitive Neuropsychology, 1988,* 5, 535-548.

Caramazza, A. (1986). On drawing inferences about the strcture of normal cognitive systems from the analysis of patterns of impaired performance: the case for single-patient studies. *Brain and Cognition, 5,* 41-66.

Case, R. (1985). Intellectual development: Birth to adulthood. Orland FL:Academic Press.

Casey, J.E., Ferguson, G.G., Kimura, D., & Hachinski, V.C. (1989). Neuropsychological improvement versus practice effect following unilateral cartid endartectomy in patients without stroke. *Journal of Clinical and Experimental Neuropsychology, 11,* 461-470.

Catron, D.W., & Thompson, C.C. (1979). Test-retest gains in WAIS scores after four retest intervals. *Journal of Clinical Psychology, 35,* 352-357.

Chelune, G.J., & Thompson, L.L. (1987). Evaluation of the general sensitivity of the Wisconsin Card Sorting Test amound younger and older children. *Developmental Neuropsychology, 3,* 81-90.

Crossen, J.R., & Wiens, A.N. (1994). Comparison of the Auditory-Verbal Learning Test (AVLT) and California Verbal Learning Test (CVLT) in a sample of normal subjects. *Journal of Clinical and Experimental Neuropsychology, 16,* 190-194.

Culbertson, W.C. & Zillmer, E.A. (1998). The construct validity of the Tower of LondonDX as a measure of the executive functions of ADHD children. *Assessment, 5,* 215-226.

Davis, S.J.C., Coltheart, M. (1999). Rehabilitation of topographical disorientation: an experimental single-case study. *Neuropsychological Rehabilitation, 9,* 1-30.

Denckla, M. & Rudel, R. (1974). Rapid automatic naming of pictured objects, colors, letters, and numbers by normal children. *Cortex, 10,* 186-202.

Denckla, M.B. (1994). Measurement of executive function, in G.Reid Lyon: *Frames of reference for the assessment of learning disabilities. New views on measurement issues.* Baltimore, Paul H.Brookes Publishing Co. (pp. 117-142).

DeRenzi, A. & Vignolo, L.A. (1962). Token test: a sensitive test to detect receptive disturbances in aphasics. *Brain, 85,* 665-678.

Dikmen, S., Machamer, J., Temkin, N., & McLean, A. (1990). Neuropsychological recovery in patients with moderate to severe head injury: 2 year follow-up. *Journal of Clinical and Experimental Neuropsychology, 14,* 507-519.

Dodrill, C.B. & Troupin, A.S. (1975). Effects of repeated administration of a comprehensive neuropsychological battery among chronic epileptics. *Journal of Nervous and Mental Disease. 161,* 185-190.

Doehring, D.J. (1968). *Pattern of impairment in specific reading disability.* Bloomington, IN: Indiana Univer. Press.

Dyche, G.M., & Johnson, D.A. (1991). Effect of repeated administration of a comprehensive neuropsychological battery among chronic epileptics. *Journal of Nervous and Mental Disease 161,* 185-190.

Evans, J.J., Emslie, H., & Wilson, B.A. (1998). External cueing systems in the rehabilitation of executive impairments of action. *Journal of International Neuropsychological Society, 1998,* 399-408.

Feuerstein, R. (1980). *Instrumental Enrichment. An intervention program for cognitive modifiability.* Illinois: Scott, Foresman & company.

Fodor, J.A. (1983). *The modularity of mind. An essay on faculty psychology.* Cambridge, Mass, MIT Press.

Gardner,M.F. (1982). *Test of Visuo-Perceptual Skills.* Washington, Special Child Publications.

German, D.J. (1993). *Word finding intervention programs.* Tuscon: Communication Skill Builders.

Gnys, J.A., & Willis, W.G. (1991). Validation of executive function tasks with young children. *Developmental Neuropsychology, 7,* 487 - 501.

Gordon, W. (1987) Methodological considerations in cognitive remediation. In M.J. Meier, A.L. Benton, & L. Diller (Eds.) *Neuropsychological rehabilitation.* Avon: Churchill Livingstone.

Heaton, R.K., Chelune, G.J., Talley, J.L., Kay, G.G. & Curtiss, G. (1993*). Wisconsin card sorting test manual. Revised and expanded.* Odessa, FL.: Psychological assessment resources.

Hersen, M., & Barlow, D.H. (1984). *Single-case experimental designs: Strategies for studying behavioral change (3rd. ed).* New York: Academic Press.

Humprey, M.M. (1982). Children's avoidance of environmental, simple task internal and complex task internal distractor. *Child Development, 53,* 736-745

Hänninen, R., Kaukovalta, E.& Kuikka, P. (1985). *Afasiakuntotus. L.S. Tsvetkovan yksilöllisen afasiakuntoutuksen menetelmiä.* Helsinki: Psykologien Kustannus.

Jöreskog, K.G., & Sörbom. D. (1993*). Lisrel 8: Structural equation modelling with SIMPLIS command language.* Chicago: Scientific Software International.

Kaplan, E. (1988). A process approach to neuropsychological assessment. In T.Boll, & B.K.Bryant. *Clinical Neuropsychology and Brain Function: Research, measurement and practice. The Master lecture series, Vol 7.* (pp. 127-167) Washington DC: APA.

Kaplan, E., Goodglass, H., & Weintraub, S. (1983*). Boston Naming Test.* Philadelphia PA. Lea & Febiger.

Karat, J. (1982). A model of problem solving with incomplete constraint knowledge. *Cognitive Psychology, 14,* 538-559.

Kaufman, A.S. & Kaufman, A.L. (1983). *K-ABC Interpretive Manual.* American Guidance Service. Circle Pines. MN.

Kavale, K. & Dobbins, D. (1993). The equivocal nature of special education interventions. *Early Child Development and Care. 86,* 23-37.

Kavale, K. (1990a). Variances and verities in learning disability interventions. In T. Schruggs & B. Wong (eds.), *Intervention research in learning disabilities* (pp. 3-33). New York: Springer-Verlag.

Kavale, K. (1990b). Effectiveness of special education. In T.Gutkin & C Reynold (eds.) *The handbook of school psychology* (pp. 868-898). New York, Wiley.

Kazdin, A. (1982*). Single-case research desings: Methods for clinical and applied settings.* New York: Oxford Press.

Kazdin, A.E. (1997). A model for developing effective treatments: progression and interplay of theory, research and practice. *Journal of Clinical Child Psychology, 26,* 114-129.

Kiresuk, T.J., Smith, A. & Cardillo, J.E. (1994). *Goal Attainment Scaling: Applications, theory, and measurement.* New York: Lawrence Erlbaum.

Klahr D., & Robinson, M. (1981). Formal assessment of problem solving and planning processes in preschool children. *Cognitive Psychology, 13,* 113-148.

Krikorian, R., & Bartok, J.A. (1998). Developmental data for the Porteus Maze test. *The Clinical Neuropsychologist, 12,* 305-310.

Krikorian, R., Bartok, J., & Gay, N. (1994). Tower of London procedure: A standard method and developmental data. *Journal of Clinical and Experimental*

Neuropsychology, 16, 840 - 850.

Leon-Carrion, J., Morales, M., Porastero, P., Dominiquez-Morales, M., Murillo, F., Jimenez-Baco, R., Gordon, P. (1991). The computerized Tower of Hanoi: A new form of administration and suggestions for interpretation. *Perceptual and Motor Skills, 73,* 63-66.

Levin, H.S., Culhane, K.A., Hartman, J., Evankovich, K., Mattson, A.J., Harward, H.,Ringholtz, G., Ewing-Cobbs, L., & Fletcher, J.M. (1991). Developmental changes in performance on tests of purported frontal lobe functioning. *Developmental Neuropsychology, 7,* 377-395.

Levin, H.S., Ewing-Cobb, L., & Fletcher, J.M. (1989). Neurobehavioral outcome of mild head injury in children. In Levin, H.S., Eisengerg H.M., and Benton, A.L. (Eds.*) Mild Head Injury.* Oxford Univer. Press, New York (pp. 189-213).

Levin, H.S., Fletcher, J.M., Kufera, J.A., Harward, H., Lilly, M.A., Mendelsohn, D., Bruce, D., & Eisenberg, H.M. (1996). Dimensions of cognition measured by the Tower of London and other cognitive tasks in head-injured children and adolescents. *Developmental Neuropsyhcology, 12,* 17-34.

MacNeill Horton, A. Jr. (1992). Neuropsychological practice effets x age: a brief note. *Perceptual and Motor Skills, 75,* 257-258.

McCaffrey R.J., Ortega, A., Orsillo, S.M., Haase, R.F., & McCoy, G.C. (1992a). Neuropsychological and physical side effects of metoprolol in essential hypertensives. *Neuropsychology, 6,* 225-238.

McCaffrey R.J., Ortega, A., Orsillo, S.M., Nelles, W.B., & Haase, R.F. (1992b). Practice effects in repeated neuropsychological assessments. *The Clinical Neuropsychologist, 6,* 32-42.

McCaffrey, R.J., Cousins, J.P., Westervelt, H.J., Marynowicz, M., Remick, S.C., Szebenyi, S., Wagle, W.A., Bottomley, P.A., Hardy, C.J., & Haase, R.F. (1995). Practice effects with the NIMH AIDS Abbreviated Neuropsychological Battery. *Archives of Clinical Neuropsychology, 10,* 241-250.

McCaffrey, R.J., Duff, K., & Westervelt, H.J. (2000). *Practitioner's Guide to Evaluating Change with Neuropsychological Assessment Instruments.* NY: Plenum Publishers.

McCaffrey, R.J., Ortega, A, & Haase, R.F. (1993). Effects of repeated neuropsychological assessments. *Archives of Clinical Neuropsychology, 8,* 519-524.

Miller, P.H., & Weiss, M.G. (1981). Children's attention allocation, understanding of attention and performance on the incidental learning task. *Child Development, 52,* 1183 - 1190.

Miller, P.H., & Weiss, M.G. (1982). Children and adult's knowledge about what variables affect selective attention. *Child Development, 53,* 543-549.

Mitrushina, M., & Satz, P. (1991). Effect of repeated administration of a neuropsychological battery in the elderly. *Journal of Clinical Psychology, 47,* 790-801.

Neyens, L.G.J., & Aldenkamp, A.P. (1996). Stability of cognitive measures in children of average ability. *Child Neuropsychology, 2,* 161-170.

Niilo Mäki Institute (1992). *Neuropsychological and achievement tests: local normative data for Niilo Mäki Institute-Test Battery.* Jyväskylä, Finland:Author.

Ozonoff, S. (1995). Reliability and validity of the Wisconsin Card Sorting Test in studies of autism. *Neuropsychology, 9,* 491-500.

Parker, E.S., Eaton, E.M., Whipple, S.C, Heseltine, P.N.R. & Bridge, T.P. (1995).

University of Southern California Repeatable Episodic Memory Test. *Journal of Clinical and Experimental Neuropsyhology. 17,* 926-936.

Passler, M.A., Isaac, W., & Hynd, G.W. (1985). Neuropsychological development of behavior attributed to frontal lobe functioning in children. *Developmental Neuropsyhcology, 1,* 349-370.

Paulsen, K., & Johnson, M. (1980). Impulsivity: A multidimensional concept with developmental aspects. *Journal of Abnormal Child Psychology, 8,* 269-277.

Pennington, B.F., Bennetto, L., McAleer, O., & Roberts, R.J.Jr. (1996). Executive functions and working memory: Theoretical and measurement issues. In G.R. Lyon & N.A.Krasnegor: *Attention, Memory and Executive Function.* Baltimore, ML.

Pennington, B.F., Groisser, D., & Welsh, M.C. (1993). Contrasting deficits in attention deficit hyperactivity disorder versus reading disability. *Developmental Psychology, 29,* 511-590.

Porteus S.D. (1965). *The Maze Test and Clinical Psychology.* Palo Alto, CA: Pacific Books.

Rapport, L.J., Brooke-Brines, D, Axelrod, B.N., & Theisen, M.E. (1997). Full scale IQ as mediator of practice effects: the rich get richer. *The Clinical Neuropsychologist, 11,* 375-380.

Raven J.C. (1984). *Manual for the Coloured Progressive Matrices (Revised).* Windsor, UK:NEFR-Nelson.

Rawlings, D.B., & Crewe, N.M. (1992). Test-retest practice effects and test score changes of the WAIS-R in recovering traumatically brain-injured survivors. *The Clinical Neuropsychologist, 6,* 415-430.

Reitan, R.M. & Davidson, L.A. (Eds.)(1974). *Clinical neuropsychology: Current status and applications.* Washington, DC: Winston.

Reitan, R.M. & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation.* Tuscon, AZ: Neuropsychology Press.

Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique. *Archives of Psychology, 28,* 286-340.

Roberts R.R.Jr., & Pennington, B.F. (1996). An interactive framework for examining prefrontal cognitive processes. *Developmental Neuropsychology, 12,* 105-126.

Robertson, I. (1997). The rehabilitation of visuospatial, visuoperceptual and apraxic disorders. In Greenwood, R., Barnes, M.P., McMillan, T.M., & Ward, C.D. (Eds.) *Neurological Rehabilitation.* Psychology Press, East Sussex, UK.

Rourke, B.P., & Gates, R.D. (1980). *Underlining Test: Preliminary norms.* Windsor, Ontario: Authors.

Rourke, B.P., & Pertauskas, R.J. (1977). *Underlining Test (Revised).* Windsor, Ontario: Authors.

Ruff, R.M., Light, R.H., & Evans, R.W. (1987). The Ruff Figural Fluency Test: A normative study with adults. *Developmental Neuropsychology, 3,* 37-51.

Schloss, P.J., Misra, A. & Smith, M.R. (1992). The use of single subject designs in applied settings. *Advances in Learning and Behavioral Disabilities, 1992*, 7, 249-290.

Schuerger, J.M., & Witt, A.C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology, 45,* 294-302.

Seron, X. (1997). Effectiveness and specificity in neuropsychological therapies: A

cognitive point of view. *Aphasiology, 11,* 105-123.

Shallice, T. (1982). Specific impairment of planning. *Philosophical Transactions of the Royal Society of London, B, 298,* 199-209.

Shatz, M.W. (1981). WAIS practice effects in clinical neuropsychology. *Journal of Clinical Neuropsychology, 3,* 171-179.

Simon, H.A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology, 7,* 268-288.

Speer, D.C. (1992). Clinically significant change: Jacobson and Truax revisited. *Journal of Consulting and Clinical Psychology, 60,* 402-408.

Spitz, H.A., Minsky, S.K., & Bessillieu, C.L. (1985). Influence of planning time and first move strategy on Tower of Hanoi problem solving performance of mentally retarded young adults and nonretarded children. *American Journal of Mental Deficiency, 90,* 46-56.

Sternberg, R.J. (1981). Intelligence and nonentrenchment. *Journal of Educational Psychology, 73,* 1-16.

Tate, R.L. Percides, M., & Maggiotto, S. (1998). Stability of the Wisconsin Card Sorting Test and the determination of reliability of change in scores. *The Clinical Neuropsychologist, 12*, 348-357.

Temple, C., Carney, R., & Mullarkey, S. (1996). Frontal lobe function and executive skill in children with Turner's syndrome. *Developmental Neuropsychology, 12,* 343-363.

Temple. C. M. (1998). *Developmental Cognitive Neuropsychology.* Psychology Press, Hove, UK.

Torgesen, J.K. (1994). Issues in the assessment of executive function. An information-processing pespective. In G.R. Lyon (ed.) *Frames of reference for the assessment of learning disabilities* (143-162). Baltimore: Brookes.

Trexler, L., & Thomas, J. (1992). Research design in neuropsychological rehabilitation. In N. von Steinbuchel, D. von Cramon, & E. Pöppel (eds.) *Neuropsychological Rehabilitation* (pp. 79-87). New York: Springer-Verlag.

Tuma, J.M., & Appelbaum, A.S. (1980). Reliability and practice effects of WISC-R estimates in a normal population. *Educational and Psychological Measurement, 40,* 671-678.

Uchiyama, C.L., D'Elia, L.F., Dellinger, A.M., Selnes, O.A., Becker, J.T., Wesch, J.E., Bai Bai Chen, Satz, P., van Gorp, W., & Miller, E.N. (1994). Longitudinal comparison of alternate versions of the Symbol Digit Modalities Tests: Comparability and moderating demographic variables. *The Clinical Neuropsychologist, 8,* 209-218.

Vik, P., & Ruff, R.R. (1988). Children's figural fluency performance: Development of strategy use. *Developmental Neuropsychology, 4,* 63-74.

Vlietstra, A.G. (1982). Children's responses to task instruction: Age changes and training effects. *Child Development, 53,* 534-542.

Wechsler, D. (1974). *Manual for the Wechsler intelligence scale for children - revised.* San Antonio, Texas:Psychological Corporation.

Wechsler, D. (1981). *Wechsler Adult Intelligense Scale – Revised Manual.* New York: The Psychological Corporation.

Weinberger, D.R., Berman, K.F., Gold, J., & Goldberg, T. (1994). Neural mechanisms of future oriented processes: In vivo physiological studies of humans. In M.M.Haith, J.Benson, R.Roberts, & B.F.Pennington (Eds.): *The Development of Future-oriented Processes.* Chicago: Universiy of Chicago Press.

Welsh, M.C., & Pennington, B.F. (1988). Assessing frontal lobe functioning in children: Views from developmental psychology. *Developmental Neuropsychology, 4,* 199 – 230.

Welsh, M.C., Cicerello, A., Cuneo, K., & Brennan, M. (1994). Error and temporal patterns in Tower of Hanoi per formance: cognitive mechanisms and individual differences. *The Journal of General Psychology, 122,* 69-81.

Welsh, M.C., Groisser, D.B., & Pennington, B. F. (1988). A normative-developmental study of measures hyphothesized to tap prefrontal functional. *Journal of Clinical and Experimental Neuropsychology, 9,* 28 [Abstract].

Welsh, M.C., Pennington, B.F., & Groisser, D.B. (1991). A normative-developmental study of executive function: a window on prefrontal function in children. *Developmental Neuropsychology, 7,* 131-149.

Welsh. M.C. (1991). Rule-guided behaviour and self-monitoring on the Tower of Hanoi disk-transfer task. *Cognitive Development, 6,* 59-76.

Zurif, E.B., Gardner, H., & Brownell, H.H. (1989). The case against the case against group studies. *Brain and Cognition, 10,* 237-255.

I

Repeated Asessment of the Tower of Hanoi Test: Reliability and Age Effects

by

Jaana Ahonniska, Timo Ahonen, Tuija Aro, Asko Tolvanen & Heikki Lyytinen
Assessment 7, 297-310

https://doi.org/10.1177/107319110000700308

II

Suggestions for Revised Scoring of the Tower of Hanoi Test

by

Jaana Ahonniska, Timo Ahonen, Tuija Aro & Heikki Lyytinen
Assessment 7, 311-320

https://doi.org/10.1177/107319110000700309

III

Practice Effects of Visuo-Motor and Problem-Solving Tests in Children

by

Jaana Ahonniska, Timo Ahonen, Tuija Aro, Asko Tolvanen & Heikki Lyytinen
Perceptual and Motor Skills (in press)

https://doi.org/10.2466/pms.2001.92.2.479

IV

Effective or Not Effective? Interpreting the Results of Intervention in Single-Case Research Design

by

Jaana Ahonniska, Timo Ahonen, Tuija Aro & Heikki Lyytinen
Manuscript (submitted), 2000

EFFECTIVE OR NOT EFFECTIVE? INTERPRETING THE RESULTS OF INTERVENTION IN SINGLE-CASE RESEARCH DESIGN

Jaana Ahonniska, Timo Ahonen, Tuija Aro, & Heikki Lyytinen

Niilo Mäki Institute,
Department of Psychology,
University of Jyväskylä,
Jyväskylä, Finland

The purpose of this article is to compare various methods of assessing change in single-case design. Results of an intervention study are presented, showing the data of pre- and post-measurements, time-series measurement, and questionnaires evaluating behavioral changes. The time-series data of the experimental group are compared to the time-series data of the control group which received no treatment. The results show that the data of pre-and post-measurement and time series measurement might be contradictory. In some cases, improvement indicated by the time series data also vanishes if compared to the spontaneous improvement resulting from repeated assessment. Additionally, the behavioral changes reported in the rating scales give different information from all the other methods. Implications of these various methods is discussed regarding reliable interpretation of intervention research.

Neuropsychological rehabilitation research has two major goals: to develop a variety of effective intervention methods, and to understand how, why, for whom and in which conditions the particular intervention methods are effective. In order to achieve these goals, one needs precise cognitive analyses of neuropsychological disorders and manifest disabilities (e.g., Taylor, Fletcher, & Satz, 1984; Teeter, 1997), specification of the mechanisms of recovery (Moehle, Rasmussen, & Fitzhugh-Bell, 1986), and analyses of processes through which treatment creates change in the targeted cognitive processes (Kazdin, 1997). Additionally, developing effective intervention requires understanding the role of modifying factors (e.g., family, school, social relations, motivation, comorbidity; Taylor, Fletcher, & Satz, 1984) and considering them in the intervention program (Ahonen, Luotoniemi, Nokelainen, Savelius, & Tasola, 1994). Last, but not least, there is a great need to develop reliable assessment tools for evaluating the efficacy of the intervention methods (Kazdin, 1997). Without sensitive, valid and reliable assessment of change, the efficacy of the intervention methods can not be evaluated, which limits gaining knowledge about the critical features of the effective interventions and restrains development of effective intervention methods. In this article, several kinds of data measured in single-case intervention research are presented and compared with each other and their value in evaluating the efficacy of the intervention is discussed.

Traditionally, demonstration of intervention effects is based on group studies. The group design, although it has its assets with normal population, has several disadvantages when applied to clinical population. Appropriate selection of the participants, matching and randomizing them in order to generate homogeneous groups is seldom possible (e.g., Seron, 1997). Additionally, dividing participants into intervention and control groups may create ethical problems in clinical settings. Furthermore, applying group research design to clinical work is laborious and the data is rarely achieved as a side product of a clinical enterprise, while routinely performed clinical intervention can provide enough data for a single case research. From a cognitive perspective it can not be assumed that all patients with similar symptoms would be suffering from same deficits or would be similarly influenced by the treatment (Caramazza, 1986). Thus, if the intervention research wants to explain which intervention works for whom and why, single-case design is commonly preferable (Kazdin, 1997).

Time series measurement applied in single-case design can use various paradigms in assessing the change (e.g., reversal, multiple baseline, and alternating treatment designs, Barlow, & Hersen, 1985). The commonly used dependent variable for time series analysis is an observable, accountable element of behavior, such as frequency of shouting, task relevant questions, or self-injurious behavior. Behavior observation has advantages as a measurement unit: it requires minimal intrusion and change in the participant's life, and repeated assessment does not result easily into practice effects or test fatigue. Nevertheless, a great amount of theoretically and clinically relevant change resulting from neuropsychological intervention, can not be assessed by behavior observation, but requires other measures such as intervention related tasks and neuropsychological or achievement tests are needed.

However, most of the neuropsychological and achievement tests are not suitable for repeated assessment because they do not have parallel or alternative forms. Even if a test has an alternative form, only very few exceptions (e.g., Parker, Eaton, Shipple, Heseltine, & Bridge, 1995) have more than one. Additionally, practice effects have been demonstrated in some neuropsychological tests even if the test has alternative forms, i.e., repeating the same structure with different contents does not prevent practice effects (Ahonniska, Ahonen, Aro, Tolvanen, & Lyytinen, in press). Various tests have also demonstrated different amounts of practice effects (Ahonniska & al., in press). In order to conclude whether the improvement of the performance reflected in time series data demonstrates genuine intervention effect or is the result of practice effects, the data of the time series measurement should be compared to a growth curve of the same tests carried out without intervention.

In this study, we present and compare different methods for demonstrating intervention effects in single case research. The data includes pre-post measurements, time series measurement based on a multiple baseline across participants design, comparison of time series measurement to the growth curve of a control group receiving no intervention, and behavior questionnaires filled out by parents and teachers during the intervention. The purpose of this experiment is to compare several kinds of data collected in single case research to analyze their value and usefulness in assessing the efficacy of intervention.

Additionally, we wanted to assess procedure specificity and generalization of treatment effects by measuring performance in tasks related to intervention material with varying degrees. Procedure specificity of intervention method indicates whether change in performance is restricted to the material or to cognitive functions trained in the intervention (Seron, 1997). Generalization of treatment effects investigates procedure specificity from a slightly different viewpoint. The lowest level of generalization indicates transfer of treatment effects from one training session to another or to alternative forms of training material. The second level of generalization can be assessed with tests that measure cognitive functions which are closely related to the intervention method. The third level of generalization indicates transfer effects of treatment which can be measured in day-to-day functioning (e.g. see Gordon, 1987).

In this research, the time series data of the intervention-related tasks reveals whether the subjects improved their performance in the tasks similar to the intervention material (no-transfer). The time series data of tasks closely related to the target area of intervention (visuo-perceptual functioning, problem solving), reveals the amount of near-transfer, and the pre-post measurement and behavior evaluations serve in assessment of far-transfer effects of intervention. The procedure specificity of the intervention was assessed by comparing the results of no-transfer tasks with tasks, which were not included to the intervention material.

Method

Participants

The participants were two girls, Mira, 9.1 years old at the beginning of the research, and Anni, 9.2 years old at the beginning of the research. Mira had a right hemisphere brain injury resulting from pre-term birth, and Anni had a fetal alcohol syndrome. Both of the participants had deficits in visuo- motor functions and problem solving as can be seen in the results of the neuropsychological and intelligence tests of Figure 1. Their language functions were relatively well developed and both of them were able to read fluently. At the beginning of the research, they were in the third grade of a special school for children with neurological and motor disorders. The IQ of Mira was 63 and the IQ of Anni was 86, measured by the Wechsler Intelligence Scale for Children- Revised (1974). Both of the children came from middle-class families.

The control group (n = 22) for time series intervention was not compared with the participants according to their age (average age 7.7 years), but according to their performance in the tests of Coloured Progressive Matrices (CPT, Raven, 1965), and the Underlining test (UL, Rourke & Gates, 1980). At the beginning of the research, the average score of CPT of the control group was 24.1 (SD 6.9). The CPT result of Anni was 23 points, and the score of Mira 16 points. The UL result of the control group was 29.8 (6.5), of Anni 36 points and of Mira 18 points. The participants of the control group were pupils of a normal primary school in Central Finland.

Intervention method

The Instrumental Enrichment method (Feuerstein, 1980) was used as a framework for the intervention. The intervention material included the nine pages of Dots and the whole workbook of Orientation in Space I, as well as 3-5 pages of both Illustrations and Comparisons. Intervention was given

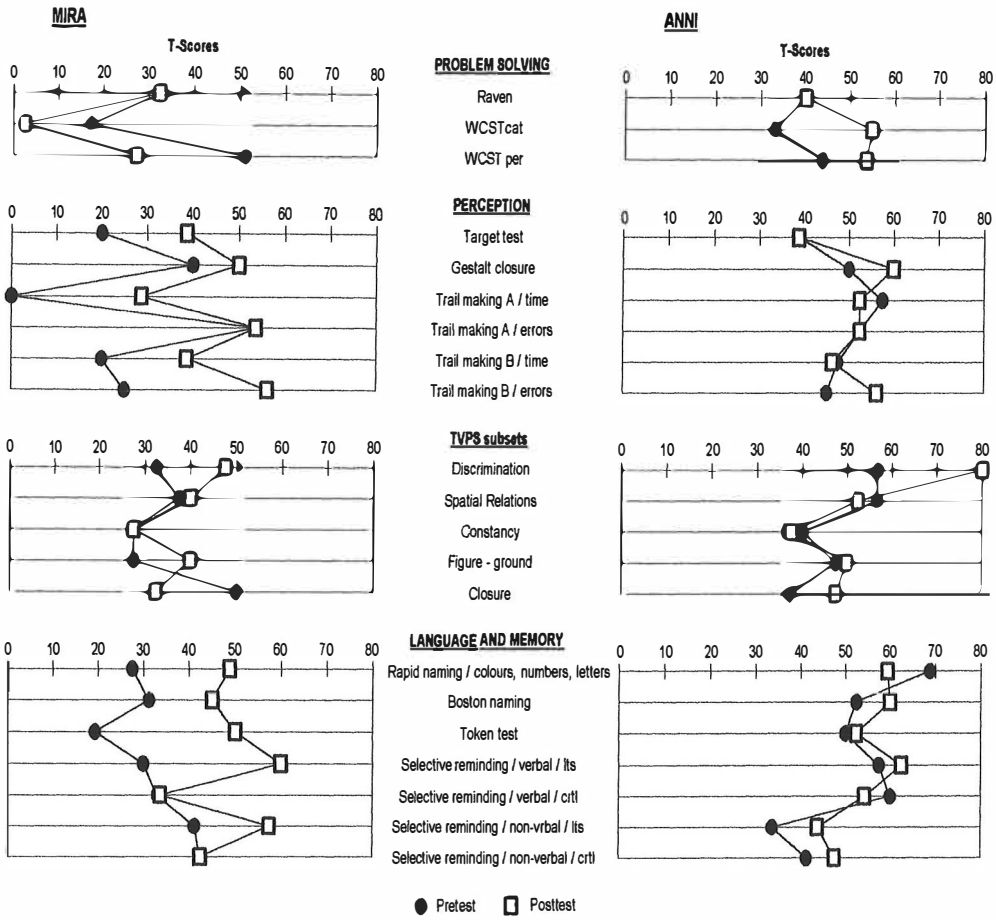


FIGURE 1. The results of the pre-post measurement of the neuropsychological tests.The tests: The Progressive Matrices (CPT, Raven, 1965), the Wisconsin Card Sorting Test/categories and perseveration (Heaton, Chelune, Talley, Kay, & Curtiss, 1981), the Target Test (Reitan & Davidson, 1974), the Gestalt Closure (Kaufman, 1983), the Trail Making Test (TMT, Reitan & Wolfson, 1992), the Test of Visual Perceptual Skills/subtests: visual discrimination, spatial relations, form constancy, figure-ground discrimination, visual closure , (TVPS, Gardner, 1982), the Rapid Naming Test (Denckla & Rudel, 1976), the Boston Naming Test (Kaplan, Goodglass & Weintraub, 1978), the Token Test (DeRenzi & Vignolo, 1962), the Selective verbal memory (Buscke & Fuld, 1974), the Selective non-verbal memory (Fletcher, 1985).

twice a week for 45 minutes in a group of two children with one psychologist as the therapist. The treatment period was 12 months (two and a half semesters) and the total amount of intervention was 60 hours. The aim of the intervention was to improve visuo-perceptual and visuo-motor functioning of the participants as well as executive functions (e.g., concentration on the task, planning, inhibition of impulsivity, evaluation of the finished task).

Measures

Pre-post measurement included a variety of neuropsychological tests, see

Figure 1. The pretest was performed one month before the beginning of the intervention and the posttest, one month after the end of the intervention.

Time series measurements were performed every two months. Mira had three and Anni five baseline measurements before the intervention. Altogether, Mira was assessed nine times and Anni twelve times.

Time series analysis included following tests:

The Tower of Hanoi (TOH, Shallice, 1982; Borys et al., 1982 Welsh et al., 1991), the Porteus Mazes Test (PMT, Porteus, 1965), the Developmental Test of Visual-Motor Integration (VMI, Beery, 1989), and the Underlining Test (UL, Rourke & Gates, 1980). Additionally, three tests, the Dots, the Orientation, and the Comparisons, were created on the basis of the intervention material and used in the time series measurement.

The Tower of Hanoi Test. The TOH test is a disk-transfer task (e.g., Borys, Spitz, & Dorans, 1982; Klahr, & Robinson, 1981; Shallice, 1982) in which the participant is supposed to transform the initial state of disk configuration to the goal state. The test included five tasks of three disks, one of four-, five-, and six-move tasks and two seven-move tasks. The dependent variable reflecting the quality of planning was counted by giving the highest possible score (amount of the minimum moves minus one) to the participant who successfully solved a particular task in the first and second trial consecutively; one point less was given if they solved the task in the second and third trial etc. Thus, the highest score possible was 24 points. The task had three alternative versions in order to diminish practice effects. A detailed description of the task procedure used in this experiment can be found in Ahonniska, & al. (a, in press). The TOH test was assumed to measure near-transfer effects of the intervention method.

The Porteus Mazes Test. The PMT (Porteus, 1965) consists of a series of increasingly difficult mazes which are designed to measure successful planning, inhibition of impulses and ability to change set. The PMT has three versions (Vineland, Extension, and Advanced) which are not parallel, but already designed to diminish the practice effect in repeated assessment. The Extension version is more difficult than the Vineland, but easier than the Advanced version. The Vineland version was used at the first, fourth and seventh assessments, the Extension at the second, fifth and eighth, and the Advanced at the third, sixth and ninth assessments. The minimum score was 7 and maximum 17 for each of the versions. The PMT was used as a measure of near-transfer effects of intervention.

The Underlining Test. The UL test (Doehring, 1968, Rourke & Gates, 1980; Rourke & Petrauskas, 1977) assesses speed and accuracy of visual discrimination with various kinds of verbal and nonverbal visual stimuli presented in single units and in combination. In this research, three subtests of the UL test were used. They were subtest 1 (Single number), subtest 7 (Two letters), and subtest 8 (Sequence of geometric forms). Out of each subtest three alternative versions were created by changing the letter or number to be searched and its location in the row of stimulus, or by changing the location and the content of the geometric sequence. The dependent variable consisted of the cumulative net score of all the three subtests (correct items minus errors). The maximum score was 117. The UL test was used as a measure of near-transfer effects of intervention.

The Developmental Test of Visual-Motor Integration. The VMI test (Beery, 1989) did not have alternative

forms, but was repeated according to the instructions of the manual at every assessment. The maximum score was 50. The test was assumed to measure near-transfer effects of the intervention method.

The Dots and the Orientation. These two tests were tasks requiring no-transfer. They were based on the intervention material and were used for assessing whether the participants learned to solve tasks directly related to the intervention material. The maximum score of the Dots test was 26 and the maximum score of the Orientation test was 12.

The Comparison. The material for the Comparison test was taken from the intervention material which was not used in this intervention. The task was used to assess whether improvement of performance was restricted to the intervention-related tasks or whether it was visible in unrelated tasks as well. The maximum score of this task was 21.

Teachers and parents filled out *questionnaires* evaluating the behavior of the participants once each semester, that is, once in four months. Altogether, four behavior evaluations were filled out for Mira and five for Anni. Mira had one and Anni two behavior evaluations filled before the beginning of the treatment.

Parents and teachers filled out the Teacher's Self-Control Rating Scale (Humphrey, 1982). A total sum score was calculated for the analysis. Minimum score was 15, and maximum score was 75, indicating good cognitive and behavioral self-control.

Additionally, the teachers filled out the Academic Performance Rating Scale (Barkley, 1991) and the Deveraux Elementary School Behavior Rating Scale (Spivack & Swift, 1982). The total scale of the Academic Performance Rating Scale of Barkley was used. The total scale includes the scales of Learning Ability, Impulse Control, Academic Performance, and Social Withdrawal. The range of the values could vary between 19 and 95, the maximum score reflecting academically successful behavior.

The Deveraux Elementary School Behavior Rating Scale does not have a total score, but has several subscales. In the current experiment, scales of peer co-operation (range of values 2-14), work organization (4-26), perseverance (2-12), impatience (4-22), irrelevant thinking (4-20), and inattention (4-24) were used in order to assess changes of behavior during the intervention. The maximum of the first three scales showed desirable behavior, while the maximum of the three last scales showed undesirable behavior. All the behavior questionnaires as well as the neuropsychological pre-post measurement were used to assess far-transfer effects of the intervention.

Results

The pre-post measurements of WISC-R revealed that Mira improved her verbal intelligence quotient (VIQ) from 68 to 74 and her performance intelligence quotient (PIQ) from 63 to 70. Anni did not show any improvement in the VIQ (from 96 to 97), but improved her performance IQ from 78 to 87. Mira showed improvement in most of the perceptual tests as well as in the tests measuring language and memory. A few subtests of the TVPS also showed some improvement. No improvement was found in the problem solving tests. Anni showed improvement in the problem solving tasks, motor performance, and two of the subtests of the TVPS. The perceptual, language and memory tests did not show improvement, see Figure 1.

Thus, Mira showed improvement in most of the pre-post assessments which were used for assessing far-transfer effects, while

Anni showed more modest gains which were limited to part of the functions trained in the intervention.

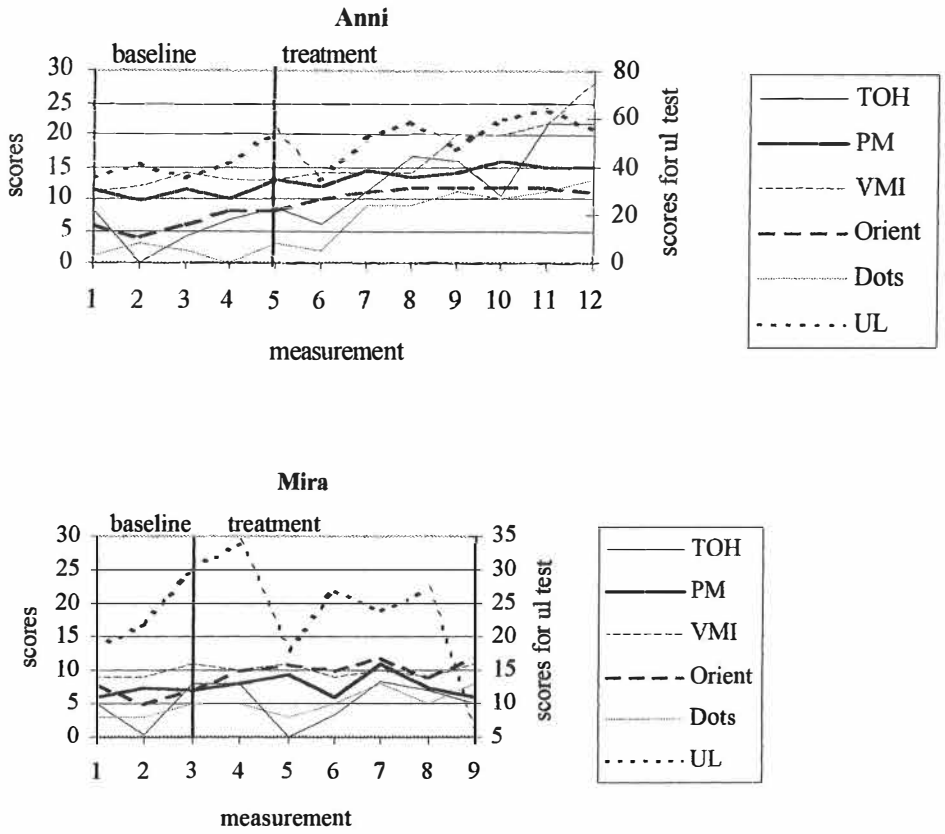


FIGURE 2. The results of the time series measurement. The tests: the Tower of Hanoi, the Porteus Mazes test, the VMI test, the Underlining test, the Dots, the Orientation and the Comparisons

Detailed observation of the time series data revealed that Mira improved her performance remarkably in only one of the no-transfer tasks (the Orientation). In the other no-transfer task (the Dots) there occurred steady but modest improvement throughout the assessments, but no improvement was seen in the near-transfer tasks. Anni showed steady improvement in one of the no-transfer tasks (the Orientation) and in two of the near-transfer tasks (the PM and the UL), and remarkable improvement in the other no-transfer task (the Dots) and in two of the near-transfer tasks (the TOH, the VMI), see Figure 2.

When the results of the time series measurements of the participants were compared with the results of the control participants, it was revealed that Mira had improved her performance in the no-transfer tasks at about the same rate as the control participants, see Figure 3.Mira's performance in the near-transfer tasks was lower than

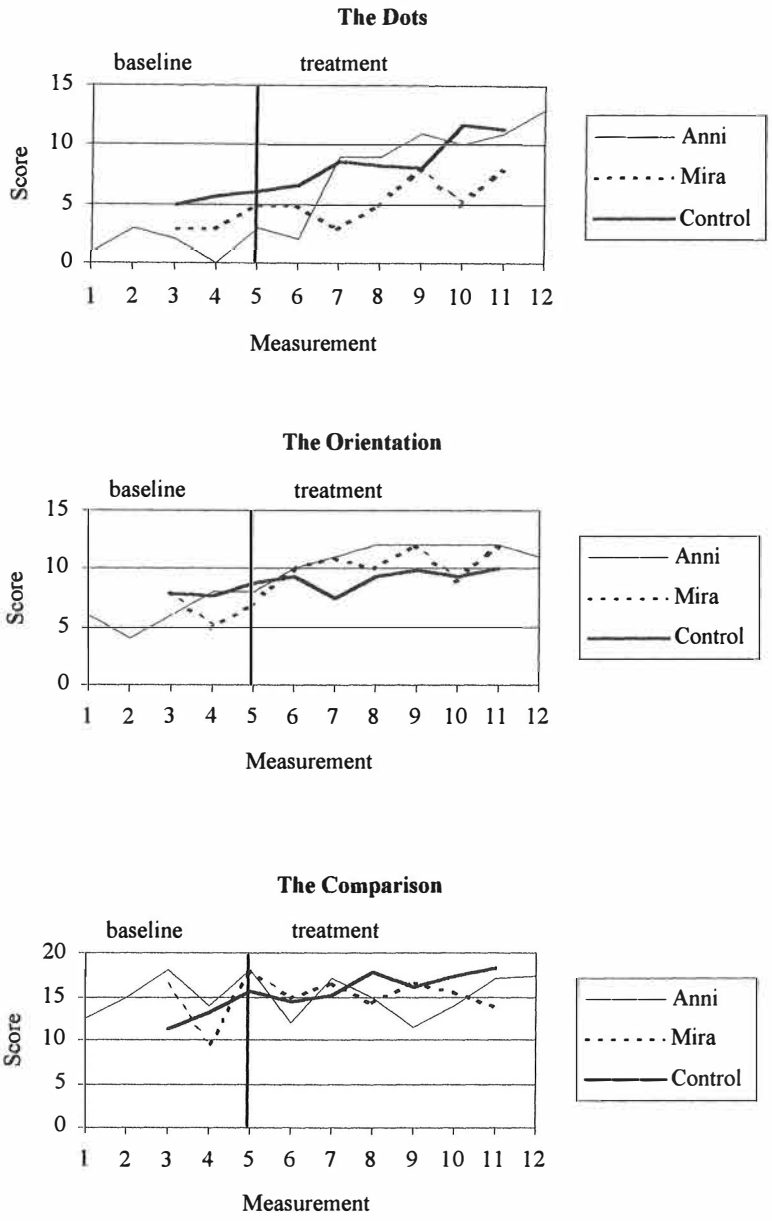
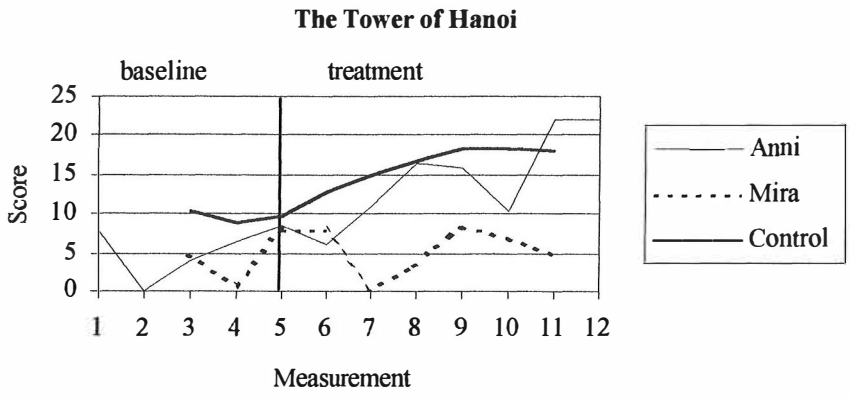


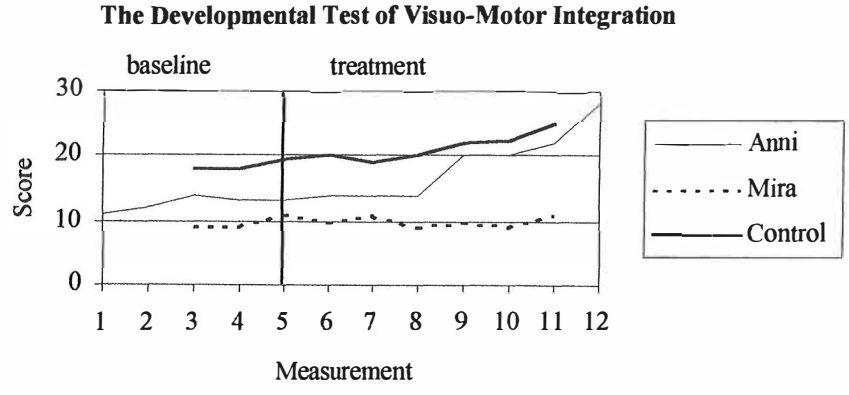
FIGURE 3. Comparison of the time series data of the participants receiving intervention with the control group without intervention. The tests: The Dots, the Orientation, and the Comparisons.

that of the control participants already from the beginning, and she improved her performance slower than the controls, see Figure 4. Anni improved her performance more than the control group in the VMI test and in the two tasks requiring no-transfer, but the improvement in other three tests was at the same level with the controls.

When the performance of the participants was compared with the control group in a task, which was not part of the intervention (the Comparisons), the participants performed at the same level with the control group.

The behavior questionnaires filled out by the teacher and parents were used for measuring the far-transfer of practice effects to daily activities. In the questionnaires, see Figure 5, Mira showed increase of desirable





(Figure 4 continues)

(Figure 4 continues)

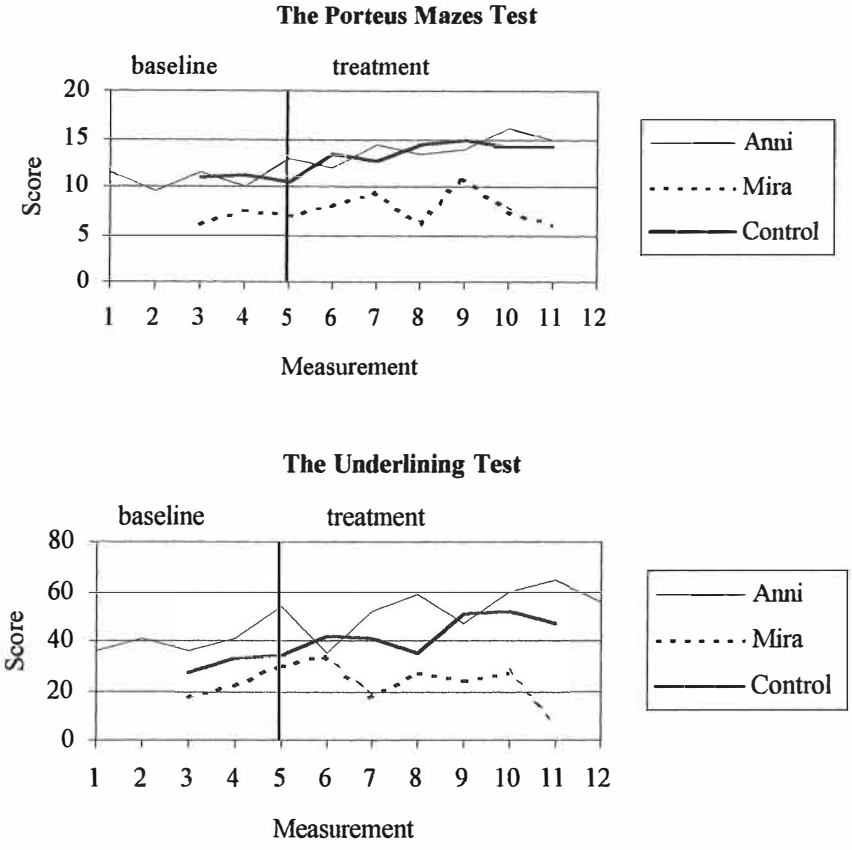


FIGURE 4. Comparison of the time series data of the participants receiving intervention with the control group without intervention. The tests (near-transfer tasks): The TOH test, the PM test, the VMI test and the UL test

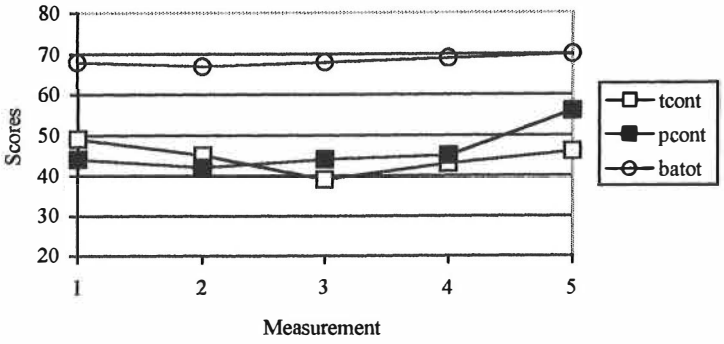
behavior and decrease of negative behavior. The teacher ratings showed almost no improvement in the behavior of Anni, and the amount of undesirable behavior even increased in some scales. The behavior evaluations of the parents showed some improvement.

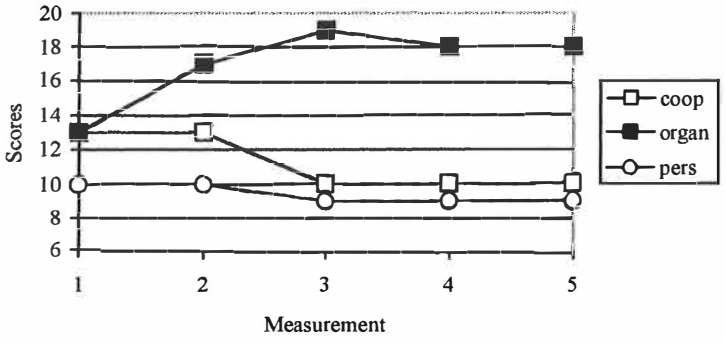
Discussion

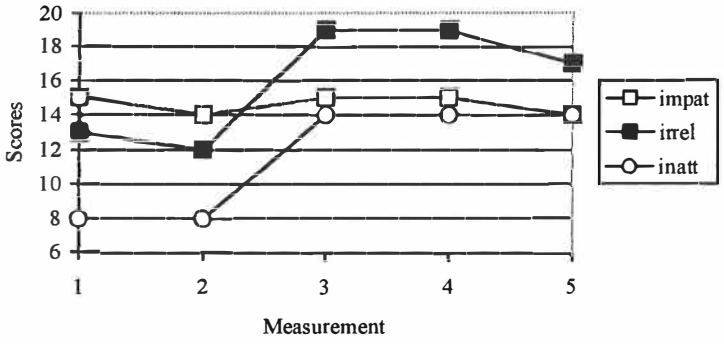
The results of this experiment show that collecting and analyzing only one kind of data in single case research can lead to biased conclusions. Mira improved her performance in most of the far-transfer tasks, including the WISC-R, most of the neuropsychological tests, and in the behavior evaluations of the teacher and the parents. Additionally, her time series data of no-transfer tasks showed some, relatively modest improvement, but the near-transfer tasks failed to show any improvement.

The results of Anni seemed to be opposite. She showed almost no improvement in the neuropsychological tests requiring far-transfer effects of the intervention. Her time series data, however, showed improvement both in the no-transfer and in the near-transfer tasks. The behavior questionnaires

Anni







(Figure 5 continues)

(Figure 5 continues)

Mira

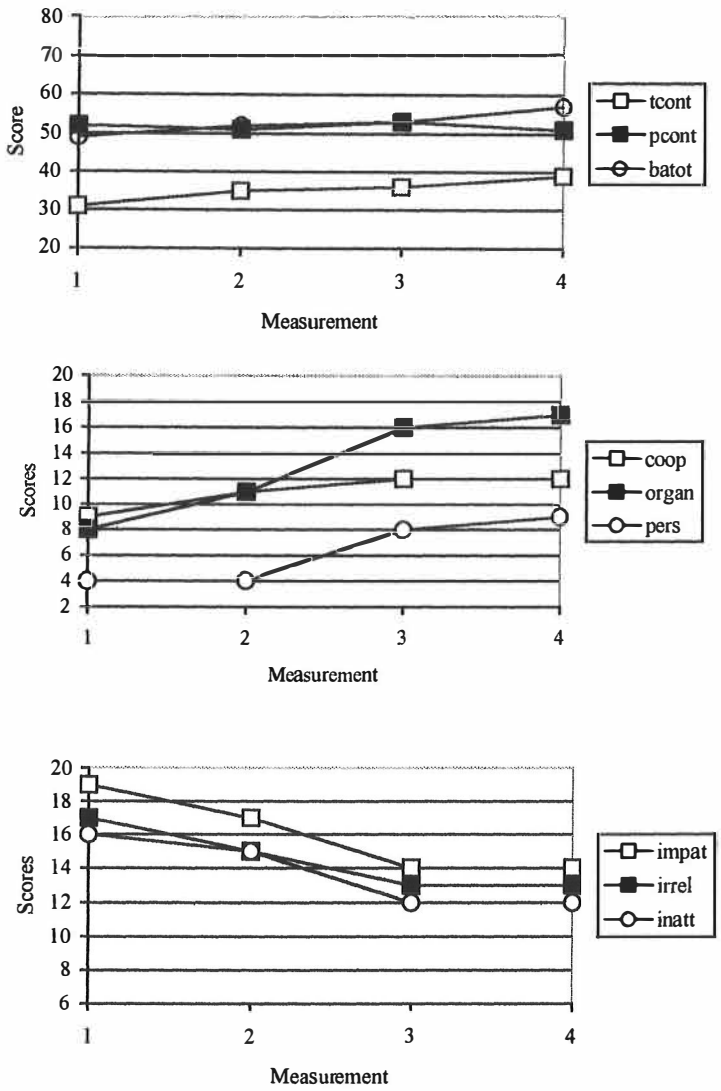


FIGURE 5. Results of the behavior questionnaires.The scales: the Teacher's Self-Control Rating Scale (filled by teacher), the Teacher's Self-Control Rating Scale (parents average), the Academic Performance Rating Scale (teacher), the subscales of co-operation, work organization, perserevance, impatience, irrelevant thinking, inattention from the Deveraux Elementary School Behavior Rating Scale (filled by teacher). The maximum in the first six scales indicates desirable behavior, and the maximum in the last three scales indicates undesirable behavior.

revealed an interesting difference between the assessments of parents and teacher. Parents found some improvement in her behavior, but the teacher reported either no meaningful

change or increase in disruptive behavior during the intervention.

How should these various, partly contradicting findings be interpreted? Which of them more reliably evaluates the intervention effects? In the single case design, the results of pre-post assessment, although commonly used, are the most unreliable. The pre- post measurement generally consists of only two measurement sessions, which causes the data to be strongly influenced by changes in mood, attention, health, familiarity with the situation and other matters considered as measurement errors. In the time series analysis these kinds of measurement errors are easier noticeable as variations of the general profile. Thus, the time series data could demonstrate the intervention effects more reliably than the pre-post data.

This experiment does not provide enough evidence to prove that the improvement of Mira's performance in the pre-post assessment resulted from intervention effects. There are two reasons for this conclusion. First, there was only slight improvement of performance in the time series data, and even that was restricted to the no-transfer tasks. Secondly, improvement in the neuropsychological pre-post assessment occurred evenly in almost all the tests. If intervention produces positive far-transfer effects without near-transfer effects, the reason for improvement could be a halo effect, or regression to mean (Speer, 1992) because her pre-intervention results were low. The reason for improved performance also could be improved motivation, improved frustration tolerance, positive attitude towards the clinician or tasks, or other unspecific effects of intervention or it could be a genuine improvement of performance, which was completely unrelated to the intervention. Anni's opposite results showing improvement in the time series measurement, but no improvement in the pre-post measurement could be interpreted as evidence of near-transfer effects, and lack of far-transfer effects of intervention. This conclusion can be made in spite of the improvement in the performance IQ, because this score is known to be sensitive to practice effects (McCaffrey, Ortega, Orsillo, Nelles, & Haase, 1992; Rapport, Brooke Brines, Axelrod, & Theisen, 1997).

Procedure specificity of the treatment (Seron, 1997) was assessed by comparing the time series data of no-transfer tasks with an intervention unrelated task (the Comparison). Improvement occurred also in the Comparison test, but the participants did not improve their performance more than the control group. Thus, this result could be an indication of a procedure specific effect of the treatment at least regarding the no-transfer tasks.

The traditional analysis of the time series data does not take practice effects into account. When neuropsychological or achievement tests are used with short test-retest intervals, practice effects could account for a great amount of improvement even if alternative versions of the tasks are used (Ahonniska et al., in press.). Thus, in order to interpret the results of the time series reliably, the time series data of the participants was compared with the time series data of a control group which did not receive intervention. In this comparison, the intervention effects seemed remarkably smaller than in the traditional analysis. In Mira's case, only the no-transfer tasks showed improvement at the same level as the control groups. In all the other tasks, she benefited from the repeated testing less than the control group and the gap between her performance and the performance of the control group widened during the intervention. In Anni's case, the traditional analysis of the time series data showed improved performance in six of the tests. When

the time series data were compared with the control data, only two of the no-transfer tasks and one of the near-transfer tasks (the VMI) showed greater improvement than the control group.

Out of all the tests used in the time series measurement, only the VMI test did not show practice effects in the control group (Ahonniska & al., in press). The improvement of performance in the control group resulted from development. The steady improvement of the performance of Anni in the first eight measurements probably shows the rate of her development in visuo-motor coordination. The fast improvement in her performance during the four last assessments is far above the normal rate of development and probably results from near-transfer effects of the intervention. Although Anni shows impressive improvement in the TOH test, the control group also shows significant practice effects in this test. Thus, the amount of improvement displayed by Anni is not remarkably greater than the possible practice effect and cannot reliably be counted as an intervention effect.

It can be argued that the results of the time series of neurologically impaired participants should not be compared to the growth curve of the normal participants. Generally, neurologically impaired subjects benefit from repeated exposure to the same test less than the normal participants (Rapport, Brooke-Brines, Axelrod, & Theisen, 1997; Shatz, 1981). Thus, the amount of change which with the normal subjects would reflect practice effects could be a significant treatment effect with participants having learning disabilities. The time series data of Mira might be evidence of this phenomenon. Her performance in the beginning of the assessment was relatively low, and after the intervention only no-transfer tasks showed improvement on the same level as the control group. In the near-transfer tests, her performance stayed on a lower level than that of the control group, and the gap between her and the control group even widened, in spite of the intervention. Thus, it could be assumed that the improvement of performance in no-transfer tasks which follows the practice effects of control group indicates significant intervention effects, and the performance in other tests indicates lack of near-transfer effects of intervention.

However, very little is known about practice effects in time series measurement, and almost nothing about the practice effects among neurologically impaired participants. Although the neurologically impaired participants possibly show fewer practice effects than the normal participants, this assumption might easily lead to underestimating the influence of practice effects and to being overly optimistic in regard to intervention effects. Thus, if psychological tests are used as an assessment tool in time series measurement, separating intervention effects from practice effects requires growth curves of various control and patient groups (see also Denckla, 1994). There is also a great need to define, "what 'leaps forward', or what qualitative, non-age referenced changes are to be accepted as therapeutic triumphs or even modest gains" (Denckla, 1994, p. 139).

The data of the behavior evaluations gave another interesting perspective to assessment of intervention effects. Based on the contradictory results between Mira's improved performance in the pre-post measurement and almost no improvement in the time series analysis, we could not conclude whether Mira improved her performance unrelated to the treatment, or whether the

improvement of the pre-post assessment resulted from measurement errors. However, both the teacher and the parents reported positive changes in Mira's behavior This supports a contention that the improvement between pre-post assessments resulted from genuine improvement of Mira's cognitive functioning and performance which either developed as an unspecific effect of intervention or was totally unrelated to the intervention.

In Anni's case, the time series data and the parents' behavior evaluation could be interpreted as a far-transfer effects of intervention. The teacher, however, reported either no change or even increase of negative behavior. This might result from the lack of generalization of the intervention effects, as there is no improvement in the pre-post assessment. The discrepancy between the parents' evaluation and teacher's evaluation might also be explained by Anni's negative attitude towards learning and the somewhat distant relationship between Anni and the teacher. Therefore, the reliability of behavior evaluations as a measurement of intervention effects might be limited by various emotional, interpersonal, or motivational reasons.

As a conclusion, time series analysis seems to be the most reliable method for assessing change. However, time series data might also provide overly optimistic results because of practice effects. Thus, if psychological or achievement tests are used in time series analysis with short assessment intervals, it is recommended to create growth curves of the same tests without intervention, using a suitable control group, and to compare the intervention results with the growth curves. Pre-post measurements of neuropsychological tests have a limited importance in children as supporting or not supporting the results of time-series analysis.

Behavior questionnaires could provide interesting information about the generalization of intervention effects to daily life. However, such factors as emotional reasons and interpersonal relationships might account for fluctuations in behavior evaluations more than any intervention. For the time being, there are no validated behavior questionnaires for all neuropsychological problems, and the ones that exist are not very useful in assessing improvement of skills such as reading. Possibly, the change of assessment of intervention effects should be shifted more towards detailed task or process analysis or towards goal achieving analysis (e.g., Kiresuk, Smith, & Cardillo, 1994) and away from neuropsychological tests. This shift would diminish the amount of practice effects, and enable the use of tests requiring less generalization of the intervention effects in order to demonstrate efficiency.

References

Ahonen, T., Luotoniemi, A., Nokelainen, K., Savelius, A., & Tasola, S. (1994) Multimodal intervention in children with attention-deficit hyperactivity disorder. *European Journal of Special Needs Education, 9,* 168-181.

Ahonniska, J., Ahonen, T., Aro, T., Tolvanen, A., Lyytinen, H. Practice effects of visuo-motor and problem solving tests in children. (in press).

Aro, T., & Ahonen, T. (1999) Tutkiva ammattikäytännön kehittäminen. In Ahonen, T. & Aro, T. (Eds.) *Oppimisvaikeudet: Kuntoutus ja opetus yksilöllisen kehityksen tukena.*

Barlow, D., & Hersen, M. (1985*). Single case experimental designs: Strategies for studying behavior change. 2nd ed.* New York: Pergamon Press.

Borys, S.V., Spitz, H.H., & Dorans, B.A. (1982). Tower of Hanoi performance of retarded young adults and nonretarded children as a function of solution length and goal state. *Journal of Experimental Child Psychology, 33,* 87-110.

Bushke H., & Fuld, P.A. (1974). Evaluating storage, retention, and retrieval in disordered memory and learning. *Neurology, 11,* 1019-1025.

Caplan, D. (1988) On the role of group studies in neuropsychological and pathopsychological research. *Cognitive Neuropsychology, 5,* 535-548.

Caramazza, A. (1986). On drawing inferences about the strcture of normal cognitive systems from the analysis of patterns of impaired performance: the case for single-patient studies. *Brain and Cognition, 5,* 41-66.

Denckla, M. (1994). Measurement of executive function. In G. Reid Lyon: *Frames of reference for the assessment of learning disabilities. New views on measurement issues.* Paul H. Brooks Publishing, Baltimore.

Denckla, M. & Rudel, R. (1974). Rapid automatic naming of pictured objects, colors, letters, and numbers by normal children. *Cortex, 10,* 186-202.

DeRenzi, A. & Vignolo, L.A. (1962). Token test: a sensitive test to detect receptive disturbances in aphasics. *Brain, 85,* 665-678.

Feuerstein, R. (1980). *Instrumental Enrichment. An intervention program for cognitive modifiability.* Illinois: Scott, Foresman & company.

Gardner, M.F. (Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *Boston Naming Test.* Philadelphia PA. Lea & Febiger.

Kaufman, A.S. & Kaufman, A.L. (1983). *K-ABC Interpretive Manual.* American Guidance Service. Circle Pines. MN.

Kazdin, A.E. (1997). A model for developing effective treatments: progression and interplay of theory, research and practice. *Journal of Clinical Child Psychology, 26,* 114-129.

Kiresuk, T.J., Smith, A. & Cardillo, J.E. (1994) (Eds.) *Goal Attainment Scaling: Applications, theory, and measurement.* New York: Lawrence Erlbaum.

Moehle, K.A., Rasmussen, J,L. & Fitzhugh-Bell, K.B. (1986). Neuropsychological theories and cognitive rehabilitation. In M. Williams & C. Long (eds.) *The rehabilitation of cognitive disabilities* (pp. 57-76) New York, Plenum Press.

Parker, E.S., Eaton, E.M., Whipple, S.C, Heseltine, P.N.R. & Bridge, T.P. (1995). University of Southern California Repeatable Episodic Memory Test. *Journal of Clinical and Experimental Neuropsyhology. 17,* 926-936.

Porteus S.D. (1965). *The Maze Test and Clinical Psychology.* Palo Alto, California: Pacific Books

Raven J.C. (1984). *Manual for the Coloured Progressive Matrices (Revised).* Windsor, UK:NEFR-Nels.

Rapport, L.J., Brooke-Brines, D, Axelrod, B.N., & Theisen, M.E. (1997). Full scale IQ as mediator of practice effects: the rich get richer. *The Clinical Neuropsychologist, 11,* 375-380.

Reitan, R.M. & Davidson, L.A. (Eds.)(1974) *Clinical neuropsychology: Current status and applications.* Washington, DC: Winston.

Reitan, R.M. & Wolfson, D. (1985) *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation.* Tuscon, AZ: Neuropsychology Press.

Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique. *Archives of Psychology, 28,* 286-340.

Rourke, B.P., & Gates, R.D. (1980). *Underlining Test: Preliminary norms.* Windsor, Ontario: Authors.

Robertson, I. (1997). The rehabilitation of visuospatial, visuoper-

ceptual and apraxic disorders. In Greenwood, R., Barnes, M.P., McMillan, T.M., & Ward, C.D. (Eds.) *Neurological Rehabilitation.* Psychology Press, East Sussex, UK.

Shallice, T. (1982). Specific impairment of planning. *Philosophical Transactions of the Royal Society of London, 298,* 199-209.

Shatz, M.W. (1981). WAIS practice effects in clinical neuropsychology. *Journal of Clinical Neuropsychology, 3,* 171-179.

Speer, D.C. (1992) Clinically significant change: Jacobson and Truax revisited. *Journal of Consulting and Clinical Psychology, 60,* 402-408.

Taylor, H.G., Fletcher, J.M., & Satz, P. (1984). Neuropsyhological assessment of children. In G. Goldstein & M. Hersen (Eds.): *Handbook of psychological assessment.* New York, Wiley.

Teeter, P.A. (1997). Neurocognitive interventions for childhood and adolescent disorders: Transactional model. In C.R.Reynols & E. Fletcher-Janzen (Eds.): *Handbook of clinical child neuropsychology.* 2nd ed. Plenum Press. New York.

Wechsler, D. (1974). *Manual for the Wechsler intelligence scale for children-revised.* San Antonio, Texas:Psychological Corporation.