PRACTICE EFFECTS OF VISUO-MOTOR AND PROBLEM SOLVING TESTS IN CHILDREN[1]

Jaana Ahonniska*, Timo Ahonen*, Tuija Aro*, Asko Tolvanen**, and Heikki Lyytinen**

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

**Department of Psychology
University of Jyväskylä
Jyväskylä, Finland

Practice effects on a visuo-motor test (the Developmental Test of Visuo-Motor Integration), a timed visual discrimination test (the Underlining Test), and two problem solving tests (the Porteus Mazes Test, and the Tower of Hanoi Test) were analyzed. Children of two age groups (means: 7.7 and 11.6 years) were chosen in order to study the effect of age on practice effects. The tests were repeated nine times with test-retest intervals of two months. The Developmental Test of Visuo-Motor Integration showed no practice effects, while the Porteus Mazes Test, the Underlining Test and the Tower of Hanoi Test showed significant practice effects. Practice effects were larger for the older age group on all the tests, except the Developmental Test of Visuo-Motor Integration. The reliability and stability of the tests were also analyzed. The Developmental Test of Visuo-Motor Integration and the Underlining test showed good reliability, but the reliabilities of the problem-solving tasks were less satisfactory. The stability of all the tests, except the Tower of Hanoi Test, were good.

One of the major objectives of clinical psychology is to measure change. Results of treatment, progress of disease, recovery, and development are assessed for research and clinical purposes. Traditionally, the change is assessed by a control group design. However, in the case of clinical samples, it is very difficult to match the treatment groups with control groups on relevant variables such as location of deficit, functional consequences of deficit, intelligence, age, and education level. Additionally, it is often impossible to randomly select participants for representation of clinical subpopulations. Control group designs also create ethical problems by leaving some of the participants without treatment. Sometimes, valuable information regarding the individual reasons for change or the lack of it, e.g. motivation, etiology and comorbidity of deficits is lost in group designs. Thus, the single-case methodology is often the best research procedure available. However, repeated assessment commonly results in improvement in performance, even without intervention (McCaffrey, Ortega, Orsillo, Nelles, &

Haase, 1992b). To make the correct conclusions, it is very important to differentiate development or intervention effect from practice effects.

Practice effects are influenced by several factors, and among these the length of the test-retest interval is very crucial. In adults, when the test-retest interval is one year or more, no practice effects have been found (Dikmen, Machamer, Temkin, & McLean, 1990; Uchiyama et al., 1994). When a test-retest interval varies between one week and six months, significant practice effects have been reported for a wide selection of psychological tests both in adults and in children (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). The practice effect becomes stronger as the test-retest interval shortens. This relationship holds when the test-retest interval varies between one week and four months (Catron & Thompson, 1979; Schuerger & Witt, 1989). Thus, in psychological practice "retesting after two years is considered adequate to avoid test-retest effects. This practice is based on common-sense principles, rather than empirical data." (Neyens & Aldenkamp, 1996, p.161). However, especially, in intervention research, repeated testing is often needed with remarkably shorter intervals, e.g. two- to six- month intervals.

If short test-retest intervals are used in adult participants, tests show various practice effects. Timed tests, requiring an infrequently practiced response, or having a single easily conceptualized solution (e.g., the Category Test), seem to be most likely to result in significant practice effects (Dodrill & Troupin, 1975). Also, memory tests generally yield significant practice effects. McCaffrey and his colleagues (McCaffrey et al., 1992b; McCaffrey et al., 1995) repeatedly found practice effects in several memory tests: the Wechsler Memory Scale-Russell's Revision (WMS-R), the digits backwards portion of the Digit Span subtest of the Wechsler Adult Intelligence Scale-Revised (WAIS-R), and the California Verbal Learning Test. Memory tests that are especially targeted to measure learning and recall show larger practice effects than tests of attention, reaction time, and recognition memory (Youngjohn, Larrabee, & Crook, 1992).

Practice effects in adults have also been found in the Trail Making Test (part B), the Grooved Pegboard Test (McCaffrey et al., 1992a, McCaffrey et al., 1992b, McCaffrey et al., 1993), the Visual Search Test (McCaffrey et al., 1995), the Paced Auditory Serial Addition Task (Dyche & Johnson, 1991b; McCaffrey et al., 1995), and the Category Test (Dodrill & Troupin, 1975). 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). Interestingly, consistent performances without practice effects have been found on the Speech Sounds Perception Test, the Motor Steadiness Test, the Seashore Rhythm Test, Simple Auditory Reaction Time (McCaffrey, Ortega, & Haase, 1993), the Finger Oscillation Test and the Span of Attention Test (McCaffrey et al., 1992a, McCaffrey et al., 1992b). Using parallel test versions in learning of word lists prevents practice effects (Crossen & Wiens, 1994; Parker, Eaton, Whipple,

Heseltine, & Bridge, 1995). In short, a wide variety of adults' tests show practice effects, especially memory tests and performance IQ in intelligence tests. Tests measuring auditory or sensory perception, motor steadiness, reaction time, or focused alertness on a task do not show significant improvement in repeated assessment.

With short test-retest intervals, children show practice effects in the intelligence tests in a similar way to adults. Repeated measurement resulted in improved performance in four of the five performance subtests on the Wechsler Preschool and Primary Scale for Children at a test-retest interval of seven to ten days (WPPSI; Longstreth & Alcorn, 1990), and in Children's Paced Auditory Serial Addition Test at a test-retest interval of four weeks (CHIPASAT; Dyche & Johnson, 1991). In the Wechsler Intelligence Scale for Children – Revised (WISC-R) and in the WPPSI, Performance IQ shows larger practice effects than Verbal IQ at a test-retest interval of six months (Neyens & Aldenkamp, 1996; Tuma & Appelbaum, 1980). The Developmental Test of Visual- Motor Integration did not show any significant practice effect at a test-retest interval of six months, while the Stroop Test, 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 with the same test-retest interval (Neyens & Aldenkamp, 1996).

Age and intelligence of the participants influence practice effects. Among adult participants, the young and middle-aged subjects show larger practice effects than the elderly ones (MacNeill Horton, 1992; Mitrushina & Satz, 1991). In addition, the subjects with average or high intelligence, benefit more from repeated testing than the subjects with lower cognitive abilities (Rapport et al., 1997). When practice effects are studied in children, development complicates the interpretations of practice effects. Development 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 may prevent or attenuate the practice effects (Crossen & Wiens, 1994; Parker et al., 1995). However, it is complicated to design truly parallel forms for psychological tests, and especially for problem solving tasks. Presumably, repeating the same task format even in an alternative task would create practice effects in problem solving tasks (Denckla, 1994), although practice effects might be smaller than those obtained using the same task (McCaffrey et al. 1992b). Very little data has been published about the magnitude of practice effects using alternative forms of tasks.

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 participants need to be assessed more often, with test-retest intervals of several weeks. Thus, growth curves are needed to assess intervention effects in the presence of practice effects (Denckla, 1994). There are no studies reporting the influence of age on practice effects in children, but it could be hypothesized that the older subjects would show larger practice effects than the younger subjects because of their higher cognitive capacity.

In this research, a test-retest interval of two months was chosen, which is clinically appropriate in most intervention research with children. On the basis of previous results, this

test-retest interval seems to create practice effects for a variety of tests. Thus, this article 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? Based on previous experiments 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.

Method

Participants

The research was performed with two groups of children. The younger children (n = 20, 10 boys and 10 girls) were in their first grade at school at the beginning of the research (m = 7.7 years), and they finished their second grade at the end of the research (m = 9.1). The older children (n = 28, 15 boys, 13 girls) were at the fifth grade at the beginning of the research (m = 11.6), and at the end of their sixth grade at the end of the research (m = 13.0). All the participants attended a normal primary school in Central Finland. The average score on the Colored Progressive Matrices (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 for the older group (z-value 0.8). The standard scores were not significantly different in the two groups (t (28) = -.807, p = Ns.).

Procedure

The tests were administered in the context of a larger study, in which several cognitive tests were administered repeatedly. The Developmental Test of Visuo-Motor Integration and the Underlining Test were administered in a classroom setting, the Tower of Hanoi Test, and the Porteus Mazes Test individually. Two psychologists administered the tests and performed the scoring.

Measures

The Developmental Test of Visuo-Motor Integration. The Developmental Test of Visuo-Motor Integration (Beery, 1982; 1989) is the most widely used developmental test of visuo-motor coordination. The task is to copy two- and three- dimensional geometrical figures according to the model. The test has 24 model figures presented in ascending order of difficulty and the possible range of the scores is from 0 to 50. The test was administered in a classroom setting according to the instructions in the manual. Each participant was allowed to work independently on the task until she/he finished it.

The Underlining Test. The Underlining 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 Underlining 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 Porteus Mazes Test. The Porteus Mazes Test (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 Porteus Mazes Test has three versions (Vineland, Extension, and Advanced) which are not parallel, but were already designed to diminish the practice effect in repeated assessments. 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 Tower of Hanoi Test. The Tower of Hanoi Test is a disk-transfer task (e.g., Borys, Spitz, & Dorans, 1982; Klahr, & Robinson, 1981; Shallice, 1982; Simon, 1975) in which the participant is supposed to transform the initial state of disk configuration to the goal state. The task evaluates the ability to generate a multistep sequence of moves (Welsh, Cicerello, Cuneo, & Brennan, 1994; Welsh, Pennington, & Groisser, 1991), strategy selection (Simon, 1975), maintaining goals in the working memory, (Welsh, et al., 1994, Welsh, et al., 1991), impulse inhibition (Roberts & Pennington, 1996), behavior monitoring, and plan revision when necessary (Welsh, et al., 1994).

The test included five tasks of three disks, one 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 minus one; that means, three trials for a four-move task and so on. The participants had to solve each task twice consecutively in order to continue to the next, more difficult task. The dependent variable, the achieved score, was a cumulative score reflecting the quality of planning. Its value was assigned by giving the highest possible score for each task (amount of the minimum moves minus one) if the participant successfully solved a particular task in the first and second trial consecutively; one point less was given if the task was solved in the second and third trial, etc. Thus, the highest score possible was 24 points. A detailed description of the procedure used in this experiment can be found in the article by Ahonniska, Ahonen, Aro, Tolvanen, & Lyytinen, (a, in press).

The equivalence of the three alternative versions of the Underlining test and the Tower of Hanoi test was assessed in a class of fourth grade children (M = 10.2, N= 32) which was divided into three groups with each group performing three different versions of the test. The average results of the versions of either of the tests did not vary significantly.

Data analysis

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

in the 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' development effect, third assessment minus four months' development effect, etc. After this subtraction, significances of the practice effects were calculated by repeated Manova using the 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.

Results

In the Developmental Test of Visuo-Motor Integration the main effect of repetition ($F_{6,02}$ =1.25, Ns.) was not significant, but the main effect of age ($F_{1,39}$ = 76.08, $p < .001$) was significant, resulting from the wide difference in the average scores between the groups. The interaction effect of repetition and age was not significant ($F_{6,02}$ = 1.24, Ns), showing no significant difference in the practice effects between the groups (see Figure 1).

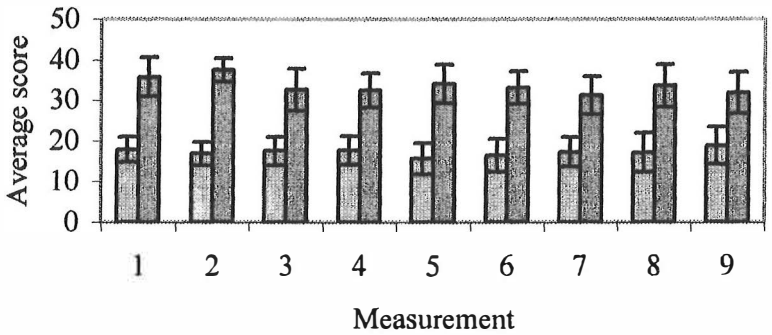

Figure 1. The development corrected results of repeated assessment of the Developmental Test of Visuo-Motor Intergration in the older (N = 28, darker column) and in the younger (N = 22, lighter column) group. Maximum Value 50. Standard Deviation displayed for each measurement.

The main effect of repetition in the Underlining Test ($F_{5,95}$ = 18.82, $p < .001$) was significant, as well as the main effect of age ($F_{1,38}$ = 126.12, $p < .001$). The practice effects were significant both in the younger ($F_{4,72}$ = 14.89, $p < .001$) and in the older group ($F_{5,02}$ = 12.50, $p < .001$), and the interaction effect of repetition and age was also significant ($F_{5,95}$ = 7.38, $p < .001$), showing that the older group showed larger practice effects than the younger group (see Figure 2).

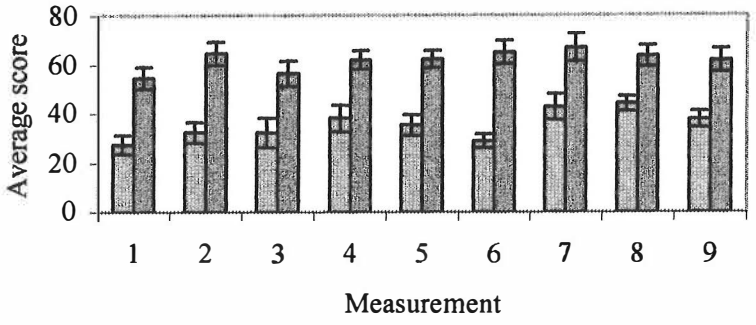

Figure 2. The development corrected results of repeated assessment of the Underlining Test in the older (N = 28, darker column) and in the younger (N = 22, lighter column) group. Maximum Value 117. Standard Deviation displayed for each measurement.

Also in the Porteus Mazes Test the main effect of repetition ($F_{4,48}$ =24.20, $p < .001$) and age ($F_{1,42} = 43.88$, $p < .001$) was significant as well as the interaction effect of repetition and age ($F_{4,48} = 10.14$, $p < .001$; see Figure 3) demonstrating that the larger practice effects in the older than in the younger group. In both of the groups the Porteus Mazes Test showed significant practice effects (younger group, $F_{4,17} = 12.04$, $p < .001$, older group, $F_{4,29} = 18.39$, $p < .001$).

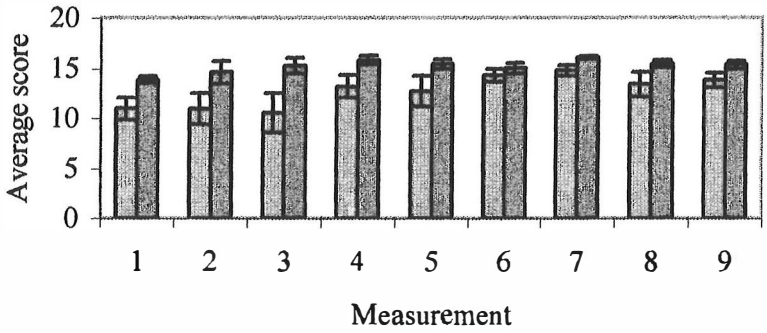

Figure 3. The development corrected results of repeated assessment of the Porteus Mazes Test in the older (N = 28, darker column) and in the younger (N = 22, lighter column) group. Maximum Value 17. Standard Deviation displayed for each measurement

The Tower of Hanoi Test showed significant main effects of repetition ($F_{5,81} = 11.94$, $P < .001$) and age ($F_{1,39} = 72.61$, $P < .001$). Significant practice effects were found in the younger group ($F_{4,99} = 4.03$, $P < .01$), and in the older group ($F_{3,46} = 14.31$, $P < .001$; see Figure 4). But significant interaction effects of repetition and age ($F_{5,81} = 2.35$, $P < .05$) showed larger practice effects in the older than in the younger group. Any of the tests did not show significant main or interaction effects for the gender variable.

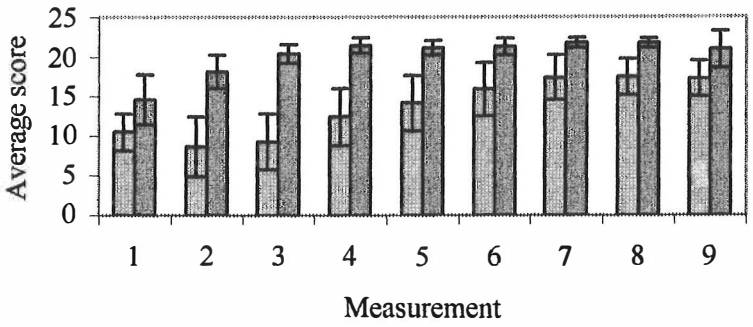

Figure 4. The development corrected results of repeated assessment of the Tower of Hanoi Test in the older (N = 28, darker column) and in the younger (N = 22, lighter column) group. Maximum Value 24. Standard Deviation displayed for each measurement.

The reliability values of the Developmental Test of Visuo-Motor Integration and the Underlining Test, measured by the squared multiple correlations for y-variable (see Note 1), were high throughout the assessments (see Table 1). Statistically, the model seemed to fit the data of the Developmental Test of Visuo-Motor Integration ($\chi^2(41) = 41.56$, $p = .45$; GFI = .78; RMSEA = 018). The model fit the data of the Underlining Test also ($\chi^2(41) = 47.13$, $p = .24$, GFI = .75; RMSEA = .06) (See Note 2). The reliability values of the Porteus Mazes Test were lower than those of the Developmental Test of Visuo-Motor Integration and the Underlining Test, and decreased slightly during the assessments. Statistically the model seemed to fit the data of the Porteus Mazes Test ($\chi^2(42) = 39.79$, $p = .57$; GFI = .80; RMSEA = .00). The reliability values of the Tower of Hanoi Test were the low in the first two assessments, but increased from the third assessment on. The statistical model fit the data: $\chi^2(20) = 13.85$, $p = .84$; GFI = .93; RMSEA = .00.

Table 1: 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 2). The stability of all the tests, except for the Tower of Hanoi Test, was high throughout the assessments. The test results of the Tower of Hanoi Test started to be relatively stable from the third assessment on.

Table 2. 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

Discussion

Repeated assessment of cognitive functions in children results easily in difficulties in interpretation of the data. The results are affected by age, by development, and by practice effects in addition to the dependent variable like rehabilitation. However, these several sources of influence are commonly not analyzed thoroughly in repeated assessment, but the development and practice effects are easily addressed as intervention effects. The aim of this experiment was to analyze the magnitude of practice effects in repeated assessment, the interaction effects of age and gender on the practice effects, and to serve as a starting point for reliable interpretation of results in repeated assessment. In previous experiments, practice effects have been studied by repeating the assessment only once or twice after the initial measurement. In the current experiment, the assessments were repeated several times, using alternative forms of most tests. This research provides growth curves for four tests. These growth curves could serve as a normative background in time series research design.

The tests differed from each other in the magnitude of practice effects. No practice effect was found in the Developmental Test of Visuo-Motor Integration, whereas the Underlining Test, the Porteus Mazes Test, and the Tower of Hanoi Test showed significant practice effects in both of the groups. In all the tests the older group performed better than the younger group. Additionally, the older subjects showed larger practice effects than the younger group on all the tests, except for the Developmental Test of Visuo-Motor Integration. The reliability and stability values of the Developmental Test of Visuo-Motor Integration and the Underlining Test were high throughout the assessments, whereas the reliability values of the Tower of Hanoi Test and of the Porteus Mazes Test were lower. The results of the Porteus Mazes Test were very stable, but the results of the

Tower of Hanoi Test started to be relatively stable and reliable only after the first two assessments.

The result of the Porteus Mazes Test, and the Tower of Hanoi Test showing practice effects, and the Developmental Test of Visuo-Motor Integration showing no practice effect, were in accordance with the earlier results (Dodrill & Troupin, 1975; McCaffrey, Ortega, & Haase, 1993; Neyens & Aldenkamp, 1996) and confirmed the hypothesis that problem solving tasks are more sensitive to practice effects than visuo-motor tasks. Even the use of alternative versions of the same test could not prevent significant practice effects either in the problem solving tasks or in the visual discrimination task. Apparently, in repeated assessment, the participants learn the basic rules, task-appropriate strategies, and plan generation; therefore changes to the task content could not prevent improvement in the performance.

While the Developmental Test of Visuo-Motor Integration showed no practice, another drawing test, the Draw-A-Person Test (Goodenough, 1926) repeated at the same intervals, showed significant practice effects in both of the groups. This interesting result could be explained by the basic difference between these two drawing tests: the Developmental Test of Visuo-Motor Integration requires copying of geometrical forms, while the Draw-A-Person Test requires drawing according to the participant's internalized image of human being. Improvement of the Draw-A-Person Test score results from increasing number of details in the human figure. Repeated assessment probably helps the participants to develop their internal image of the human figure, and to create more detailed drawings, which results in practice effects. The Developmental Test of Visuo-Motor Integration, on the contrary, has carefully controlled test material, which is not too sensitive to variations in attention, memory or general cognitive abilities. The increasing complexity of the tasks in the Developmental Test of Visuo-Motor Integration with variations in angles, directions, overlapping forms, and three-dimensionality is probably sensitive to the genuine developmental changes in the visuo-motor coordination, and improvement of the results can not be achieved only by repeated measurement.

In this experiment, one test (the Developmental Test of Visuo-Motor Integration) was administered without alternative versions, two other tests (the Underlining Test and the Tower of Hanoi Test) had alternative versions which were equally difficult, and one of the tests (the Porteus Mazes Test) had three versions of increasing difficulty. The method of increasing level of difficulty is supposed to take into account practice effect, but has not been investigated earlier. The current results show that repeating the Porteus Mazes Test resulted in a slightly staircase-shaped learning curve. Every three consecutive assessments of the Porteus Mazes Test were on the same level, this was followed by a rise in the curve between the third and fourth, and between the sixth and seventh measurement when the easiest version was administered after the most difficult one. Thus, the alternative versions of the Porteus Mazes Test can take practice effect into account, and probably some of development effect. However, the Figure 3 shows that even after correcting the data for development effect, practice effect in the older group is so large that the performance improves in the three first

assessment sessions in spite of the increasingly difficult versions.

The method of using increasingly difficult versions in repetitive assessment is very rarely used but might be a very useful method in repeated assessment. In some instances - especially in the case of problem solving tests - the practice effects can be so large that using normal alternative versions is not practical, even if growth curves for the certain test and the certain age group were be available. However, the Porteus Mazes Test provides another model in diminishing and possibly even minimizing the practice effects, which might be applicable also to other tests and help to interpret the results of the intervention.

As was expected, the older group showed more practice effect than the younger group on most of the tests. Only on the Developmental Test of Visuo-Motor Integration, where generally no practice effect was found, no difference between the groups was found. In the problem solving tasks, the difference in practice effect between the older (11-12 years) and younger (7-8 years) groups could be due to the development of executive functions occurring between these years (Passler, Isaacs, & Hynd, 1985). The older children have better working memory, (Case, 1985), are less impulsive (Humphrey, 1982; Vlietstra, 1982), have better selective attention in ignoring irrelevant stimuli (Miller & Weiss, 1981; 1982), and have better hypothesis testing and impulse control (Welsh, Groisser & Pennington, 1988). All this probably explains the differences in level of performance and practice effects between the groups.

It could be argued that our method of estimating the development effect results in bias due to the non-linear development between the ages seven and nine, and eleven and thirteen. It is possible that the development in the older group is remarkably slower than in the younger group. Thus, in the most extreme case our method of estimating the development effects could result in underestimating the development effect in the younger group and overestimating it in the older group. This would mean that the practice effect would be large in the younger group and small in the older group. However, our data showed opposite results, that the older group actually showed larger practice effects than the younger group. Possibly, the real practice effects of the younger group are even smaller and the real practice effects of the older group even larger than estimated. In any case, the direction of the conclusions seems to be appropriate.

Additionally, when the current method of calculating the practice effect was compared to the available developmental data on the Developmental Test of Visuo-Motor Integration (Beery, 1989), the Porteus Mazes Test (Krikorian, & Bartok, 1998), and the Underlining Test (Rourke, & Gates, 1980), the standardized scores gave results similar to our method. In both of the groups, the standard score of the Developmental Test of Visuo-Motor Integration did not change from the first to the last assessment in both of the groups. In the Underlining Test, 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 Porteus Mazes Test, 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, proved to be less reliable measures than the Developmental Test of Visuo-Motor Integration and the Underlining Test. The individual learning curves showed that the results of the Tower of Hanoi Test and the Porteus Mazes Test were notably variable in the first few assessments. Most of the subjects needed three assessments in order to learn the requirements of these two tasks and to stabilize their performance at a relatively high level. In addition, the decrease of reliability values of the Porteus Mazes Test in the last assessments probably results from a ceiling effect.

The stability values of the Developmental Test of Visuo-Motor Integration, the Underlining Test and the Porteus Mazes Test showed that interindividual differences were very constant from the first to the last assessment. This is explained by the wide age difference between the groups. The stability of the Tower of Hanoi Test was not satisfactory in the first few assessments, but improved after the third assessment. However, the interindividual differences in the Tower of Hanoi Test did not ever become as stable as in the other tests. This might be a result of the method of task administration. In order to diminish error variance and to improve stability, more failures should be allowed at several levels of the task before interruption. The stability might be increased also by using a four-disk version, which would then help avoiding a ceiling effect (Ahonniska, et al., b, in press).

As a summary, it could be assumed that if the score on the Developmental Test of Visuo-Motor Integration shows significant improvement in intervention research, the intervention very likely improves visuo-motor coordination. Significant increases in scores on the three other tests could result from practice effects. If we want to make reliable conclusions in intervention research, in assessing the progress of disease, and in any other experiments requiring repeated assessment, several alternative methods could be used. First, one could provide age appropriate growth curves for the relevant tests, as done in this research, and compare the results with the growth curves. Second, because problem solving tasks are particularly prone to practice effects, one could design alternative forms where each form is more difficult than the previous one, as in the Porteus Mazes Test in the current research, or one could change both the form and the content of a test (Denckla, 1994). Both of these options are difficult to design. The third alternative is to assess cognitive functions with tests which have not shown practice effects (e.g., sensory perception tests, reaction time, alternative forms of word list learning), to apply widely experimental single case methods, and to use behavioral observation.

Notes

Note 1. The traditional test-retest reliability measured by correlation analysis considers the differences between the subjects in the measured ability to be totally consistent from one measurement to another (stability equals 1.0). All the variation in these differences is considered to result from the weaknesses of the test. The LISREL analysis using simplex model to evaluate reliability allows variation in the differences between 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 session to another. Stability value indicates how constant the differences are in the measured ability between subjects.

Note 2. The Goodness of Fit Index (GFI) and the Root Mean Square Error of Approximation (RMSEA) together with χ^2 test are commonly used indexes measuring the goodness of fit between the model and the data values. The model is considered fitting, if the GFI is > .90, or the RMSEA is < .05 or the p-value of χ^2 > .05.

Authors' notes

This research was supported by the Foundation for the Haukkala's Child Psychiatric Institute, Central Finland's Foundation of the Finnish Cultural Foundation and the Support Foundation of Handicapped Children. We express our gratitude to Hanna Mäntynen for collecting the data, and the teachers and pupils of Muurame elementary school. We also thank Riva Freiman and Marilyn Schneider for correcting the language.

References

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

Ahonniska, J., Ahonen, T., Aro, T., & Lyytinen, H. (b) Suggestions for revised scoring of the Tower of Hanoi task. *Assessment* (in press)

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.

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.

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.

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.* (pp. 117-142).Baltimore: Paul H.Brookes Publishing Co.

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 Neuropsy- chology, 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. (1991a) Effect of repeated administration of a comprehensive neuropsychological battery among chronic epileptics. *Journal of Nervous and Mental Disease, 161,* 185-190.

Dyche, G.M., & Johnson, D.A. (1991b) Development and evaluation of CHIPASAT, an attention test for children: II. Test-retest reliability and practice effect for a normal sample. *Perceptual and Motor Skills, 72,* 563-572.

Goodenough, F.L.(1926) *Goodenough intelligence test.* NY:Yonkers on Hudson.

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

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

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.

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).

Longstreth, L.E., & Alcorn, M.B. (1990) Susceptibility of Wechsler Spatial Ability to experience with related games. *Educational and Psychological Measurement, 50,* 1-6.

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

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

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 assessme- nts. *The Clinical Neuropsychologist, 6,* 32-42.

McCaffrey, R.J., & Westervelt, H.J. (1995) Issues associated with repeated neuropsychological assessme- nts. *Neuropsychology Review, 5,* 203-221.

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.

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 Neuropsychology, 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.

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. (1965) *Guide to using the Coloured Progressive Matrices, set A, Ab, B.* London:H.K.Lewis.

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.

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

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

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

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

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

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

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.

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

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

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

Youngjohn, J.R., Larrabee, G.J., & Crook, T.H. III (1992) Test-retest reliability of computerized, everyday memory measures and traditional memory tests. *The Clinical Neuropsychologist, 6,* 276-286.