Author(s): Syväoja, Heidi; Tammelin, Tuija H.; Ahonen, Timo; Räsänen, Pekka; Tolvanen, Asko; Kankaanpää, Anna; Kantomaa, Marko

Title: Internal Consistency and Stability of the CANTAB Neuropsychological Test Battery in Children

Year: 2015

Version:

Please cite the original version:

Syväoja, H., Tammelin, T. H., Ahonen, T., Räsänen, P., Tolvanen, A., Kankaanpää, A., & Kantomaa, M. (2015). Internal Consistency and Stability of the CANTAB Neuropsychological Test Battery in Children. Psychological Assessment, 27(2), 698-709. https://doi.org/10.1037/a0038485

Internal consistency and stability of the CANTAB neuropsychological test battery in children

Heidi J Syväoja
LIKES – Research Center for Sport and Health Sciences, Jyväskylä, Finland and University of Jyväskylä, Finland

Tuija H Tammelin
LIKES – Research Center for Sport and Health Sciences, Jyväskylä, Finland

Timo Ahonen
University of Jyväskylä, Finland

Pekka Räsänen
Niilo Mäki Institute, Jyväskylä, Finland

Asko Tolvanen
University of Jyväskylä, Finland

Anna Kankaanpää
LIKES – Research Center for Sport and Health Sciences, Jyväskylä, Finland

Marko T Kantomaa
LIKES – Research Center for Sport and Health Sciences, Jyväskylä, Finland and Imperial College London, U.K.

Author note

Heidi J Syväoja, LIKES – Research Center for Sport and Health Sciences, Department of Psychology, University of Jyväskylä; Tuija H Tammelin, LIKES – Research Center for Sport and Health Sciences; Timo Ahonen, Department of Psychology, University of Jyväskylä; Pekka Räsänen, Niilo Mäki Institute; Asko Tolvanen, Department of Psychology, University of Jyväskylä; Anna Kankaanpää, LIKES – Research Center for Sport and Health Sciences; Marko T Kantomaa, LIKES – Research Center for Sport and Health Sciences, Department of Epidemiology and Biostatistics, MRC–HPA Centre for Environment and Health, Imperial College London.

This study was funded by Finnish Ministry of Education and Culture and the Academy of Finland (grant 273971).

The data of the present study was part of a larger research project, which aimed to determine the associations of physical activity, sedentary behavior, cognitive functions and academic achievement in elementary school-aged children: Syväoja, H.J., Tammelin, T.H., Ahonen, T., Kankaanpää, A. & Kantomaa, M.T. 2014. The Associations of Objectively Measured Physical Activity and Sedentary Time with Cognitive Functions in School-aged Children. PloS one 9 (7): e103559. This study is also part of the doctoral dissertation.

Correspondence concerning this article should be addressed to Heidi Syväoja, LIKES – Research Center for Sport and Health Sciences, Viitaniementie 15a, 40720 Jyväskylä, Finland. E-mail: heidi.syvaoja@likes.fi

Abstract

The Cambridge Neuropsychological Test Automated Battery (CANTAB) is a computer-assessed test battery widely use in different populations. The internal consistency and one-year stability of CANTAB tests were examined in school-aged children. Two hundred-thirty children (57% girls) from 5 schools in the Jyväskylä school district in Finland participated in the study in spring 2011. The children completed the following CANTAB tests: a) visual memory (Pattern Recognition Memory [PRM] and Spatial Recognition Memory [SRM]), b) executive function (Spatial Span [SSP], Stockings of Cambridge [SOC], and Intra-Extra Dimensional Set Shift [IED]), and c) attention (Reaction Time [RTI] and Rapid Visual Information Processing [RVP]). Seventy-four children participated in the follow-up measurements (64% girls) in spring 2012. Cronbach's alpha reliability coefficient was used to estimate the internal consistency of the nonhampering test, and structural equation models were applied to examine the stability of these tests. The reliability and the stability could not be determined for IED or SSP because of the nature of these tests. The internal consistency was acceptable only in the RTI task. The one-year stability was moderate-to-good for the PRM, RTI, and RVP. The SSP and IED showed a moderate correlation between the two measurement points. The SRM and the SOC tasks were not reliable or stable measures in this study population. For research purposes, we recommend using structural equation modeling to improve reliability. The results suggest that the reliability and the stability of computer-based test batteries should be confirmed in the target population before using them for clinical or research purposes.

Keywords: CANTAB, internal consistency, reliability, stability, children

Computer technology can make valuable contributions to neuropsychological assessments (Cernich, Brennana, Barker, & Bleiberg, 2007; Parsey & Schmitter-Edgecombe, 2013). Neuropsychological test batteries have been used for one-off psychological assessments, for evaluating changes over time, and for evaluating the effects of interventions (Lowe & Rabbitt, 1998).

The Cambridge Neuropsychological Test Automated Battery (CANTAB) is one of the oldest computer-based test batteries used to evaluate neurocognitive functioning, particularly in clinical trials research. CANTAB is a Windows-based program administered via a touch screen computer. CANTAB tests are mainly nonverbal and allow investigation of visual and spatial memory, executive function, working memory, planning, different aspects of attention, and other areas of cognition (Cambridge Cognition Ltd., 2006).

Previous research suggests that CANTAB is a suitable method to measure cognitive functions in 4- to 90-year-old individuals, particularly in people with Alzheimer's, Parkinson's disease, schizophrenia, and autism (Lowe & Rabbitt, 1998; Luciana, 2003). Studies have also demonstrated that it is sensitive to deficits due to several neuropsychological and psychiatric conditions, especially in the elderly (e.g., Blackwell et al., 2003; Levaux et al., 2007; Robbins et al., 1998; Sahakian & Owen, 1992), but also in children (e.g., Gau & Shang, 2010; Fried, Hirshfeld-Becker, Petty, Batchelder, & Biederman, 2012; Luciana, Lindeke, Georgieff, Mills, & Nelson, 1999; Rhodes, Riby, Matthews, & Coghill, 2011). Although studies have shown that CANTAB tests can discriminate clinical populations from normal controls, little is known about the validity of these tests compared to traditional neuropsychological tests in the general population (Smith, Need, Cirulli, Chiba-Falek, & Attix, 2013).

CANTAB tests are mainly based on traditional neuropsychological tests (Cambridge Cognition Ltd., 2006). These traditional tests have been highly used and their validity and reliability has been carefully assessed (Ahonniska, Ahonen, Aro, Tolvanen& Lyytinen, 2000; Gnys & Willis, 1991; Mammarella, Pazzaglia, & Cornoldi, 2008; Halperin, Sharma, Greenblatt, & Schwartz, 1991). However, the reliability of the CANTAB tests has been inadequately described in earlier studies (Luciana & Nelson, 2002). According to Luciana (2003), internal consistency coefficients were high (.73 – .95) in 4–12-year-old children. However, to our knowledge, there are no other studies establishing the internal consistency agreement of CANTAB tests in a child population.

Furthermore, studies measuring the test-retest reliability of CANTAB have been sparse. According to Lowe and Rabbit (1998), the test-retest agreement for the CANTAB tests in an elderly adult population was either moderate, ranging from .70 to .86, or low, ranging from .09 to .68 for four week time interval. Fisher et al. (2011) observed quite low intra-class correlations for CANTAB subtests, Spatial Span length, and Spatial Working Memory errors (ICC = .51 – .59) for a three-week interval in healthy children. However, according to Gau and Shang (2010), intra-class correlations for CANTAB tests (Intra-Extra Dimensional Set Shift, Spatial Span, Spatial Working Memory, Stockings of Cambridge) ranged from .55 to .94 for time interval of 14–42 days in a group of 10 children with Attention deficit hyperactivity disorder (ADHD). To our knowledge, there are no other studies establishing the test-retest agreement for CANTAB tests in a child population (Henry & Bettenay, 2010; Luciana, 2003), and the results of existing studies are, to some extent, inconsistent.

It is important to examine the reliability and the stability of the tests because low reliability limits both the sensitivity to diagnose clinical conditions and the sensitivity to detect

changes in cognition over time (Lowe & Rabbitt, 1998). In addition, low reliability and stability limit the usefulness of the test as a research and clinical tool. Moreover, the use of CANTAB and similar computer-based neuropsychological test batteries is increasing worldwide and that is why it is important to determine the psychometric properties of these kinds of test batteries. The purpose of this study was to evaluate the internal consistency and the one-year stability of seven CANTAB tests measuring visual memory, executive function, and attention in elementary school-aged children.

Method

Participants

The data of the present study was part of a larger research project, which aimed to determine the associations of physical activity, sedentary behavior, cognitive functions and academic achievement in elementary school-aged children. In spring 2011, 230 fifth and sixth graders (48% of 475 eligible, 57% girls, M_{age} = 12.19, SD = .63) from five schools in the Jyväskylä school district in Finland participated in the study. When the children were invited to participate in the study, they were given an information pack containing a leaflet for themselves, a letter for their parents/guardians, and a consent form. Participation in the study was voluntary, and all the participants and their parents were informed about their right to drop out of the study at any time without a specific reason. Only children with a fully completed consent form (signed by a parent/guardian and the child) on the day of the first measurements were included in the study. In 46% of families, the highest level of parental education was tertiary level education. Seventy-seven percent of the parents were married or cohabiting. Six percent of the children had a diagnosed learning difficulty. The children had normal or corrected-to-normal vision. They participated in normal curriculum-based instruction, and the language of instruction was Finnish.

During spring 2012, students who were fifth graders in spring 2011 were invited to participate in the follow-up measurements. Seventy-four children participated in these follow-up measurements (49% of 151 eligible, 64% girls, $M_{age2011}$ = 11.73, SD = .37, $M_{age2012}$ = 12.82, SD = .04). The study was performed according to the principles of the Declaration of Helsinki and the Finnish legislation and was approved by the Ethics Committee of the University of Jyväskylä.

Procedures

CANTAB (a PaceBlade Slimbook P110 tablet PC with a 12-inch touch-screen monitor and Windows XP Professional operating system, CANTABeclipse version 3) was used to assess a broad range of cognitive functions: a) visual memory (Pattern Recognition Memory [PRM] and Spatial Recognition Memory [SRM]), b) executive function (Spatial Span [SSP], Stockings of Cambridge [SOC], Intra-Extra Dimensional Set Shift [IED]), and c) attention (Reaction Time [RTI] and Rapid Visual Information Processing [RVP]) (Table 1). The tests were run individually with the help of a trained research assistant and according to the standard protocol. The standard instructions for the tests were provided in the CANTAB manual and were translated into Finnish. The execution of the tasks required about 45 minutes. The test battery was administered in a silent room without distractions. A Motor Screening Task measuring simple psychomotor speed and accuracy was used as a training procedure at the beginning of a test session.

Measures

Visual memory. Visual memory performance was assessed with PRM and SRM. PRM measures recognition memory for visual patterns in a two-choice forced discrimination paradigm. In the presentation phase, the children were presented with 12 different geometric patterns one after the other in the center of the screen. Beforehand, they were asked to remember

the patterns. In the recognition phase, the children were presented with 12 trials of two patterns: a pattern they had already seen and a novel pattern. They were asked to choose a pattern they remembered having seen. The target patterns were presented in the same order as in the first time. This subtest was repeated with a new set of the 12 patterns to be remembered. The score in this task is based on the number of correct responses (maximum 24).

SRM measures recognition memory for spatial locations in a forced-choice paradigm. In the presentation phase, the children were presented with a white square on the screen in five different locations and asked to remember the locations where they had seen the square. In the recognition phase, the children were shown two squares in different locations: one in the same location as before and the other in a new location. The children were instructed to choose the location where they remembered seeing the square. The target locations were presented in the same order as in the first time. The block of five trials was repeated four times in total. The score in this task is based on the number of correct responses (maximum 20).

Executive function. Children's executive functions were assessed with SSP, SOC and IED tests. The SSP is based on the Corsi Blocks task (Milner, 1971), which measures the length of the visuospatial memory span. In each trial, there are 10 white boxes on the screen, and the color of a specified number of boxes changes one by one. The children were directed to reproduce the sequence by touching the same boxes in the same order that the boxes changed their color. If the child reproduced the correct sequence, he/she passed to the next difficulty level, where one more box was added to the sequence. The child has three attempts at each level. If the third attempt was unsuccessful, the task was terminated. The task starts with a two-box sequence and ends with a nine-box sequence, which is the highest possible level to proceed. The score in the task is based on the length of the maximum sequence that the child can reproduce.

The SOC is a computerized version of the Tower of London task (Owen, Downes, Sahakian, Polkey, & Robbins, 1990; Shallice & Shallice, 1982) measuring spatial planning and spatial working memory. At the beginning of each problem, the children were presented with a computer screen split into two parts. In both parts of the screen, there were three vertical stockings and three colored balls in predetermined order. The children were required to move the colored balls in the lower part of the screen to the same position in the stockings as they are in the upper part of the screen. They were asked to use only a specific number of moves (two, three, four, or five) to fulfill the goal. The balls can be moved one at time by touching first the required ball and then the target location. The balls cannot be moved outside the stockings, and the lower balls cannot be moved before the upper balls. If the child took more than double the number of moves required to fulfill the goal, the task was terminated. If three consecutive problems were terminated, the entire test ended. The score in this task is based on the number of problems the child solves with the minimum number of moves.

IED is a computerized analogue of the Wisconsin Card Sorting test and measures rule acquisition and reversal in a set-sifting condition. Specifically, it measures the ability to focus attention on different stimuli within a relevant dimension and shift attention to a previously irrelevant dimension. There are nine stages, with increasing difficulty in this task. The children were instructed to choose one of the two different dimensions: one is correct, and the other is incorrect. According to immediate feedback given from the computer, they were expected to choose the correct pattern and learn the rule. The children progressed through the task by satisfying a predetermined criterion of learning the rule at each stage (six consecutive correct choices). If the children failed to learn the rule during 50 trials at any stage, the test was terminated.

The first two stages of the task measured simple discrimination (stage 1) and simple reversal (stage 2), and the children had to choose between two purple patterns. In stages 3 and 4, compound stimuli (a purple pattern and a white lined drawing) were presented. In these stages, the children had to continue to respond to the previously relevant dimension (the purple pattern) and ignore the presence of the new irrelevant dimension (the white lined drawing) (nonoverlapping compound discrimination [stage 3] and compound overlapping discrimination [stage 4]). These stages were followed by compound reversal (stage 5). In stage 6, the first attentional shift is required (the intradimensional shift). The children were presented with new compound stimuli: novel shapes of each of the two dimensions (the purple pattern and the white line drawing). They had to continue to respond to the relevant dimension (the purple pattern). This stage was followed by the intradimensional reversal (stage 7). In stage 8, the compound stimuli changed again, but this time the children had to shift their attention (the extradimensional shift) and respond to the previously irrelevant dimension (the white lined drawing). Stage 9 involves extradimensional reversal. The score in this task is based on the number of stages completed.

Attention. The children's attention abilities were assessed with RTI and RVP. RTI measures the children's speed of response to an unpredictable visual target. In the unpredictable condition, the yellow spot appears in any of five circles on the screen. The children were instructed to hold down the press pad button until they saw the yellow spot and then touch the middle of the correct circle as quickly as possible. The children took part in rehearsal trials and 15 task trials. The scores in this task are based on reaction time (ms) and movement time (ms).

The RVP measures the sustained attention and is similar to the Continuous Performance Task. There is a white box on the screen where digits from 2 to 9 appear in a pseudo-random

order at the rate of 100 digits per minute. The children were directed to touch the press pad button every time they saw the following target sequence: digits 3, 5, and 7 (the 357 mode). During the practice stage, the children received hints and feedback from the computer, which declined gradually. In the assessment stage, the children received no hints or feedback. This stage took 3 minutes. The score in this task is based on RVP A', which measures how good the child is at detecting the target sequences (range: .00 to 1.00; bad to good).

Statistical analysis

For the statistical analyses, the SPSS 19.0 for Windows statistical package (SPSS (2010) IBM SPSS Statistics 19 Core System User's Guide (SPSS Inc., Chicago, IL) and the Mplus statistical package (Version 7; Muthèn & Muthèn, 1998–2012) were used. Logarithmic transformations were applied to variables with skewed distributions, and gender differences were tested using the independent samples *t*-test. The differences between the subjects with complete data and the subjects who did not participate in the follow-up measurement were tested using the independent samples *t*-test for continuous variables and Pearson's Chi-Square test for dichotomously scored variables. Pearson's correlation coefficients were calculated for continuous variables between the different sections of each test. Effect sizes (Cohen's *d*, standardized mean difference) were calculated for repeated measures. The reliability of each nonhampering test was estimated with Cronbach's alpha reliability coefficient (α). As preliminary analysis of stability, Pearson's correlation for continuous variables and tetrachoric correlations for dichotomously scored variables were calculated between the assessments.

To examine the stability of the CANTAB tests, structural equation models were applied. To combine a very large number of measured variables for each latent factor, item parcels were constructed by summing every third pattern to same parcel (Little, Cunningham, Shahar, &

Widaman, 2002). Item parcels were used in order to achieve continuous indicators and not to end up with too large model in terms of the ratio of sample size to number of free parameters (Herzog & Boomsma, 2009; Westland, 2010). Underlying latent traits were assumed to be unidimensional. The constructed parcels were then used as indicator variables in the confirmatory factor analyses.

The measurement models were first specified at both measurement times to test the association between the observed variables and the underlying factors. After demonstrating the fit of the measurement models, longitudinal confirmatory factor analyses were performed. The baseline stability model, in which the factor (factors) in the second assessment (2012) was predicted by the factor (factors) in the previous measurement point (2011), was estimated. To detect time invariance in the latent constructs, equality constraints were imposed on the corresponding factor loadings across two time points. Furthermore, if the invariance assumption of a stability model was supported, a more parsimonious model in which all the factor loadings were fixed to be one was estimated.

Full information maximum likelihood (FIML) estimation with robust standard errors (MLR) was used under the assumption of data missing at random. Item response theory (IRT) modeling using FIML estimation was applied to the CANTAB tests with hampering nature.

The goodness-of-fit of the cross-sectional and longitudinal models was evaluated by the Satorra–Bentler scaled χ^2-test, the comparative fit index (CFI), the Tucker–Lewis Index (TLI), the root mean square error of approximation (RMSEA), and the standardized root-mean-square residual (SRMR). The model fits the data well if the *p*-value for the χ^2-test is non-significant. CFI and TLI values close to 0.95, an RMSEA value below 0.06, and an SRMR value below 0.08 indicate good fit between the model and the observed data (Hu & Bentler, 1999). A Satorra–

Bentler scaled χ^2 difference test was conducted for the nested models. If the χ^2-test produces a non-significant loss of fit for the constrained model as compared to the baseline stability model, the equality constraints are supported.

Results

In this study, Cronbach's alpha reliability coefficients were calculated for cross-sectional data in 2011 to estimate the internal consistency of the CANTAB tests. Structural equation modeling was used to estimate the one-year stability of the CANTAB tests in longitudinal data setting.

The mean values, standard deviations, and gender differences for the cross-sectional sample at the first measurement point are presented in Table 2. The performance of the boys and girls did not differ in the CANTAB test, except in the RTI, where the boys' five-choice movement time (2011: t(228) = 3.30, p = .001, 2012: t(86) = 2.81, p = .006) and reaction time (2011: t(228) = 4.07, p < .001, 2012: t(86) = 2.39, p = .020) were shorter than that of the girls. The mean values and the standard deviations for the follow-up sample and the effects sizes (Cohen's d) for the repeated measures are presented in Table 3. The subjects with complete data (n = 74) did not differ from the subjects who did not participate in the follow-up measurements (n = 156) with respect to the highest level of parental education, family structure, family income, or diagnosed learning difficulties. These groups also did not differ in their performance in the CANTAB tests.

Reliability

Cronbach's alpha reliability coefficients were .65 for the PRM number of correct responses, .21 for the SRM number of correct responses, .87 for the RTI five-choice movement

time, .66 for the RTI five-choice reaction time and .49 for the RVP A'. For hampering tests (SSP, SOC and IED), Cronbach's alpha could not be determined.

Stability

Visual memory. The distribution of the PRM number of correct responses was negatively skewed, and 50% of children made two mistakes or less. For the PRM number of correct responses, Pearson's correlation between the assessments was $r(74) = .53$ ($p < .001$). Three parcels from 24 patterns (incorrect/correct response) of PRM were constructed by summing eight items into the same parcel. These subscales were then used as the indicator variables in the confirmatory factor analyses. The baseline stability model fitted the data well. Invariance assumption was supported, with an insignificant scaled difference in the χ^2 value (χ^2 (2) = .89, p = .64). In addition, the model in which all the loadings were fixed to be one was confirmed (χ^2 (4) = 4.67, p = .32). The goodness-of-fit statistics of the more constrained model for the PRM number of correct responses were good (χ^2 (12) = 17.30, p = .14, CFI = .96, TLI = .95, RMSEA = .04, SRMR = .13). The estimation results of the model are presented in Figure 1. The stability for the PRM was .80. The factor loadings and the measurement error variances were significant.

Three parcels from 20 patterns (incorrect/correct response) were formed also for the SRM. The estimation results of the cross-sectional measurement models in 2011 and 2012 revealed that none of the parcels loaded significantly on the hypothesized factor. Therefore, a stability model for SRM could not be determined. Correlations among 20 trials of SRM were calculated, and only 25 correlations of 190 were statistically significant. The correlations were modest and ranged from -.14 to .37, with one exception: The second and fifth trials in the third block correlated with each other ($r(229) = .77$, $p < .001$). The correlation for the SRM number of

correct responses between 2011 and 2012 was $r(72) = .30$ ($p = .009$). The level of performance remained the same in 20% of the children, it improved in 43%, and it declined in 37%.

Executive function. The stability model for the SSP could not be determined, because of the nature of the test. The distribution of children's span lengths are presented in Table 4. The correlation between the SSP span length in 2011 and in 2012 was $r(72) = .37$ ($p = .001$). The span length stayed the same in 34% of children, improved in 53%, and declined in 14%.

The cross-sectional IRT models for SOC were estimated for dichotomously scored items (problem solved/not solved in minimum moves). Only 5 of the 12 items loaded significantly on the hypothesized factor in 2011 and 3 of the 12 items in 2012. According to the results of the estimated baseline stability model, there was no significant association between the latent variables measured in 2011 and 2012. The correlation for SOC problems solved in the minimum number of moves between 2011 and 2012 was $r(72) = .23$ ($p = .046$). The level of performance stayed the same in 11% of children, improved in 58%, and declined in 31%.

As the numbers of errors in stages 3 to 7 and 8 to 9 of the IED test were highly correlated with each other, a cross-sectional two-factor model of the number of errors was estimated. All the items loaded on the hypothesized factors, except for the errors in stage 7 in 2012. There was no significant correlation between the factors either in 2011 or in 2012. In addition, the results of the estimated baseline stability model revealed that the regression coefficient between the latent variables measured in consecutive years was significant only for the latent variable measured by the errors in stages 8 and 9 ($b = .68$, $SE = .08$, $p < .001$). Therefore, the stages from 3 to 7 were discarded from further analyses. For the dichotomous variables, which indicated whether the child completed the stage, tetrachoric correlations were calculated. The tetrachoric correlations between stages 8 and 9 were .99 ($p < .001$) in 2011 and .96 ($p < .001$) in 2012. The tetrachoric

correlation for stage 8 between 2011 and 2012 was .38 (p = .047). For stage 9, it was .64 (p < .001). The level of performance remained the same in 69% of the children, it improved in 19%, and it declined in 12%. In 2011, 66% of the children passed the test, and 72% passed the test in 2012.

Attention. For the RTI five-choice reaction time, Pearson's correlation between the assessments was $r(74) = .63$ ($p < .001$). For the RTI five-choice movement time, it was $r(74) = .50$ ($p < .001$). For structural equation modeling, three subscales of reaction time and movement time at both measurement points were formed from 15 patterns in the RTI (and each subscale was divided by 10). The results of the estimated cross-sectional two-factor model confirmed that the subscales of reaction time and movement time loaded on their hypothesized factors. The baseline stability model fitted the data reasonably well. The scaled χ^2-difference test produced a non-significant loss of fit for the constrained stability model compared with the baseline model ($\chi^2(4) = 6.59$, $p = .16$). In addition, a more constrained model, in which the loadings were fixed to be one, was supported, with an insignificant difference in the χ^2 value ($\chi^2(8) = 6.91$, $p = .55$). The more constrained stability model for the RTI reaction time and the movement time is presented in Figure 2. The goodness-of-fit statistics of the more constrained model for the RTI reaction time and the movement time were good (χ^2 (58) = 78.51 (58) $p = .04$, CFI = .96, TLI = .95, RMSEA = .04, SRMR = .15). The stability coefficient for the RTI movement time was .67. For the RTI reaction time, it was .78. All the factor loadings and the measurement error variances were significant.

For the RVPA', Pearson's correlation between the assessments was $r(74) = .39$ (p = .001). The RVP A' (multiplied by 10) scores of the three blocks of the test were used as indicator variables in the structural equation modeling. The distribution of the RVP A' number of correct

responses was highly negatively skewed. The baseline stability model fitted the data reasonably well. The invariance assumption was supported, with an insignificant difference in the χ^2 value (χ^2 (2) = 4.41, p = .11). When all the factor loadings were constrained to be one, a scaled χ^2 difference test produced a significant loss of fit (χ^2 (4) = 10.96, p = .03). A constrained stability model for RVP A' is presented in Figure 3. The goodness-of-fit statistics of the constrained model for RVP A' were reasonable good (χ^2 (10) = 17.94 (10) p = .06, CFI = .91, TLI = .87, RMSEA = .06, SRMR = .25. The stability coefficient for RVP A' was .62. All the factor loadings and the measurement error variances were significant.

Discussion

In this study, the internal consistency and the one-year stability of seven CANTAB tests were examined in elementary school-aged children. According to Cronbach's alpha reliability coefficients, only the RTI five-choice movement time increased above a typically accepted level of .7 (Nunnally & Bernstein, 1994). According to the structural equation modeling, the PRM number of correct responses and the RTI reaction time had high levels of stability. In addition, the RTI movement time and the RVP A' had a moderate level of stability. Furthermore, the SSP span length and the IED number of children who completed at least stage 8 seemed to have a moderate correlation between the two measurement points. The SRM and the SOC tasks were not reliable or stable measures in this study population.

Visual memory. Cronbach's alpha reliability coefficients for the PRM and the SRM were below the accepted level of .7 (Nunnally & Bernstein, 1994), indicating that the total scores of the tests should be used with caution. In addition, the patterns in the SMR did not correlate with each other, and those patterns that did showed only a modest correlation, fluctuating around

zero. This finding does not support that of Luciana (2003) who reported that the internal consistency coefficients for CANTAB tests ranged from .73 to .95 in 4–12-year-old children. To increase the reliability of the PRM, we recommend estimating the measurement errors away by using structural equation modeling in the analyses.

One possible explanation for the low internal consistency of the PRM may be the ceiling effect. According to Luciana and Nelson (2002), children reach an adult level of performance in the PRM by the age of 7 years, after which ceiling levels are reached. Thus, it may be problematic to discriminate 9- to 12-year-old children with high ability. In our study, the children were about 12 years. Generally, their performance was high in the PRM test. Fifty percent of the children made two errors or less in the test. It seems that the errors the children made were sparse and random, which may explain why the patterns of the tests do not have a high correlation with each other. However, this was not the case with the SRM because the children's performance did not reach the level of ceiling.

The stability was good for the PRM number of correct responses (.80). However, a stability model for SRM could not be determined because none of the parcels of SRM loaded significantly on the hypothesized factor. The finding regarding PRM's stability is consistent with that of Lowe and Rabbitt (1998) who reported that the test-retest correlation for PRM was .84 in a healthy adult population (ages 60–82 years).

In summary, according to the results of this study, the PRM task of CANTAB proved to be a stable measure for 12-year-olds. To increase the reliability of the PRM, we recommend using structural equation modeling in research analyses. The SRM task was not a reliable measure.

Executive function. The traditional version of Corsi Blocks test has been shown to be a reliable test for young adults (α = .85) (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001) and children (α = .79) (Mammarella et al., 2008). In this study, the CANTAB version of the Corsi blocks task, SSP, seemed to be an adequate tool for measuring visuospatial memory capacity in the children aged 12 years. There was variance in the results of the SSP among the children. In addition, most of the children improved their performance one year later, which is line with a previous study that found that children approximates functional maturity in the SSP task at the age of 12 years (Luciana & Nelson, 2002). Nevertheless, we cannot draw advanced conclusions about the reliability of the SSP because its internal consistency could not be determined. Thus, we cannot say how much is real variance and how much is measurement error.

In the case of the SSP, either stability model could be determined. However, the correlation for the SSP span length between 2011 and 2012 was .37. According to previous studies, the test-retest reliability for SSP was .51 for a three-week interval (Fisher et al., 2011), .55 for 14–42 days interval (Gau & Shang, 2010) in children and .64 for a four-week interval in the elderly (Lowe & Rabbitt, 1998).

The CANTAB test, SOC is identical to the traditional Tower of London (TOL) task (Owen et al., 1990; Shallice & Shallice, 1982). According to the results of this study, it seems that the patterns in the SOC are not informative. Other studies had also raised questions about the psychometric characteristics of traditional TOL tasks (Bishop, Aamodt-Leeper, Creswell, McGurk, & Skuse, 2001). Humes, Welsh, Retzlaff, and Cookson (1997) reported low internal consistency for the TOL (split-half reliability of .19 and Cronbach alpha of .25). In addition, according to Ahonniska et al. (2000), a similar task, the Tower of Hanoi (TOH) did not have satisfying reliability (simplex estimator) in the first two assessments, but improved with

repetition, when children aged 8 and 12 years participated in nine repeated assessments of the task during 18 months. The test-retest reliability of the traditional TOL task has been reported to be quite low: .5 for 30–40 day time interval (Bishop et al., 2001), but also acceptable: .72 for 25 min time interval in preschool children (Gnys & Willis, 1991). However, Ahonniska et al. (2000) reported relatively high stability for TOH (Beta= .82–1.00, depending on the performance index) after the second assessments. The temporal stability of the CANTAB version, SOC, has been reported to quite low: 26–.60 for four week time interval (depending on the performance index) (Lowe & Rabbitt, 1998), but also acceptable: .72 for 14–42 day time interval in children with ADHD (Gau & Shang, 2010).

Ahonniska et al. (2000) proposed that large intraindividual variation due to different rates of learning may explain the low reliability (simplex estimator) in the first few assessments. Also, Lowe and Rabbitt (1998) suggested that one possible explanation for the low temporal stability of SOC might be the novelty of the task. Performance in the tests of executive function can abruptly improve when an individual discovers an optimal strategy, but the performance improves less or not at all if a strategy is not found. The performance may even decline if an incorrect strategy is attempted. Different practice effects will weaken the test-retest reliability (Lowe & Rabbitt, 1998). According to Anderson, Anderson, and Lajoie (1996), performance in the TOL improves approximately at the age of 11 years due to a developmental spurt. In our study, 58% showed improved performance, and 31% showed a decline in performance. However, Bishop et al. (2001) reported that average scores in the retest and the initial test were nearly the same in 7–15-year-old children indicating that task novelty cannot account for the low test-retest reliability. Bishop et al. (2001) suggested that any variation due to individual

differences in neurology in executive function performance may be overwhelmed by powerful factors other than brain development influencing the performance in executive function tasks.

In addition, it has been reported that a simpler TOL (two, three, four, or five moves) gives ceiling effects in older children (Anderson et al., 1996; Krikorian, Bartok, & Gay, 1994). However, according to Luciana, Collins, Olson, and Schissel (2009) and Luciana and Nelson (2002), 11–12-year-old children do not reach adult levels of performance. In our study, the children did not reach the ceiling levels, which suggests that the ceiling effect does not explain the low reliability in this case.

IED task is based on traditional the Wisconsin Card Sorting test. In this study, the analysis showed that only stages 8 and 9 were informative, and around 70% of the children passed the whole test. According to previous studies, shifting ability improves with age and reaches ceiling levels by age 12 years (Anderson, 2002; Luciana & Nelson, 2002). According to Luciana and Nelson (1998), it is typical for normal adults to make significantly more errors in the extradimensional shift stage (at stage 8) than in earlier stages. This phenomenon was repeated in the 8-year-olds but not in the younger children (Luciana & Nelson, 1998). Substantial amount of errors in the extradimensional shift stage was also observed in the current study with 12-year-olds, and it seems that only stages 8 and 9 discriminate children with weaker performance from children with general performance.

In this study, according to the tetrachoric correlations, the minimum stability for IED stage 8 completed was 0.38. For IED stage 9 completed, it was .64. In previous studies, the temporal stability for the CANTAB IED task was reported to be .78 for 14–42 day time interval (Gau & Shang, 2010) in children, .40–.75 (depending on the performance index) in adults (Henry

& Bettenay, 2010) and .09–0.70 (depending on the performance index) for four week time interval in the elderly (Lowe & Rabbitt, 1998).

In summary, it seems that only the SSP of these executive function tests could work well in healthy 12-year-old children. The IED was quite stable, but it seemed to be too easy for the 12-year-olds and did not discriminate between children with higher abilities. In this study, the SOC was not a reliable or stable measure.

Attention. In this study, the internal consistency coefficient for the RTI five-choice movement time reached the accepted level of .7 (Nunnally & Bernstein, 1994), whereas internal consistency was below this accepted level for the RTI five-choice and RVPA'. The stability was good for the reaction time and moderate for the RTI movement time and the RVP A'. Sustained attention has been assessed with the Continuous Performance Task before, which is similar to the CANTAB RVP task. The split-half reliability for the computerized Continuous Performance Task was reported to be .38–.95 (depending on the performance index) (Conners, Epstein, Angold, & Klaric, 2003; Halperin et al., 1991). According to previous studies, the test-retest reliability ranged from .55–84 (depending on the performance index) for 4.8 month interval (Conners et al., 2003) and 0.55–.84 (depending on the performance index) for three month interval (Halperin et al., 1991).

In this study, the RVP A' performance of the children was very high. According to Halperin et al. (1991), there were no difficulties related to ceiling or floor effects in the Continuous Performance Test in 7–10-year-old boys, but they assumed that ceiling effects would occur in older children. According to the results of this study, the ceiling effect was observed in the RVP task, which may explain the low internal consistency. In addition, the low stability may also be the result of the easiness of the test. Due to high level of performance, a single random

mistake may induce the effect of varying performance between the assessments, and affect stability.

In summary, according to the results of this study, the RTI task of CANTAB is a reliable and stable measure for 12-year-olds. When using this test in research analysis, we recommend estimating the measurement errors away by using structural equation modeling to improve the reliability of the RTI. The RVP 357 mode was easy for the 12-year-olds in this study. It might be useful to utilize a more difficult version of the test with a greater number of target sequences in this age group.

Taken together, the internal consistency of most of the CANTAB tests used in this study was low, and two of the tests did not measure the phenomena they were supposed to measure with satisfying reliability. Some previous studies reported that computerized versions of traditional neuropsychological tests are not equivalent to traditional manual tests. For example, Feldstein et al. (1999) reported that computerized versions of the Wisconsin Card Sorting test were not similar to the manual version. In addition, according to Smith et al. (2013), CANTAB subtests measuring executive function, speed of processing, visual memory, and working memory correlate only modestly with traditional subtests. The reason for these findings is unknown. Perhaps, computerized test sessions are more sensitive to attentional disruptions than manual sessions, or they may not offer similar perceptual characteristics to manual versions.

In the present study, stability of the measured CANTAB tests was moderate supporting the previous studies (Fisher et al., 2011; Gau & Shang, 2010; Lowe & Rabbitt, 1998). However, in these previous studies the test-retest interval has been shorter, only few weeks, whereas, in the present study, stability was measured with one year interval. This longer time interval may affect

the stability, because it is expected that the cognitive abilities of 11-year-old children develop during one year. Therefore, measuring their performance twice in the beginning with a shorter interval between the measures would have enabled calculation of the test-retest reliability, and ruling out the developmental effects.

Lastly, there are a lot of advantages in computerized testing: computer technology has increased the efficiency, ease, and standardization of administration and saved time and money related to testing. Electronic data capture and automatic results scoring have minimized human errors in scoring and data entry and increased the accuracy of timing and response latencies. (Cernich et al., 2007; Parsey & Schmitter-Edgecombe, 2013). Other advantages of computerized technology – particularly for the measurement of attention, motor, and memory functioning – are the availability of almost unlimited alternate forms and an increased number of trials. This minimizes practice effects and allows more assessments at shorter time intervals compared to traditional measures. A large number of trials and accurate assessment of reaction times results in data that is normally distributed and on a true interval scale (Betts, Mckay, Maruff, & Anderson, 2006). This is particularly important when slight changes in performance across time are assessed.

Computer technology also provides test administration conditions that are accommodating for individuals with particular needs (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 2014). Touch-screen technology facilitates use by young children and certain clinical groups, and allows more reliable assessment of motor function and processing speed compared to traditional measures using an individual administrator wielding a stopwatch. In addition, non-verbal culture-neutral test stimuli are often used, which allow the

application of computerized test batteries (like CANTAB) for individuals from different racial, ethnic, geographic, or sociocultural backgrounds. (Luciana, 2003; Henry, 2010). CANTAB also has simple standardized test administration; thus, it is easy to use and no previous IT or scientific training is needed to set-up and administer the test (Cambridge Cognition Ltd., 2006).

Despite the potential advantages that computerized technology offers for neuropsychological testing – especially the ease of building ready algorithms for calculating indexes and scores – computer-assisted assessment also produces a risk factor. Due to commercial competitive reasons, the companies providing these tools may not always publish detailed information about these algorithms or the psychometric properties of the tasks. The general characteristics of scoring algorithms and the accuracy of the algorithms as well as technical evidence should be documented and reviewed periodically (AERA, APA & NCME, 2014). According to The Finnish Psychological Test Committee (2013), test batteries (computerized or not) that do not provide a satisfactory level of information about the psychometric properties of the tasks in the technical manual should not be used in clinical practice. Likewise, in clinical practice, an expert (a psychologist or an MD) should always do the interpretation of the test results. There are two reasons for this. First, to use CANTAB or other computerized test batteries, there needs to be a person who supervises and paces the introduction of tests and monitors the assessment process (Luciana, 2003; AERA, APA & NCME, 2014). Secondly, the interpretation and the clinical conclusions and decisions made based partly on the test results produce a juridical situation where there needs to be a responsible decision-maker. Most countries do not allow automated decision-making in healthcare.

There will be rapid growth in the usage of computer-assisted test batteries. However, little is known about the psychometric properties of computerized batteries as well as how

performance on computerized batteries correlates with traditional neuropsychological measures. In addition, it is also vital to have more information about the reliability and validity of these batteries in all clinical target populations of interest as well as in samples derived from the normal population.

Strengths and limitations

To our knowledge, this is the first study examining the internal consistency and one-year stability of the CANTAB tests in healthy 12-year-old children. This study provides valuable and important information on the psychometric characteristics of CANTAB tests in a child population. The study sample in this study in 2011 was quite large and representative of Finnish children at the age of 12. On the other hand, the number of children in the follow-up measurements in 2012 was limited, which attenuated the statistical power. However, subjects with complete data did not differ from the subjects who did not participate to follow-up measurements in socioeconomic positions or performance in the CANTAB tests. In addition, the models were estimated with FIML estimation, which uses all information available and takes missing data into account. The age-range of the children studied was narrow, which limits the application of the results to different age groups or developmental stages. In addition, the study sample was culturally homogeneous including children only from Finland. Also, we did not measure the performance of the children again immediately after the first measurement. Thus, we could not calculate intra-class correlations for the CANTAB tests. Furthermore, several measurement points during the year would give more accurate information on the effects of practice on the performance in the tests.

Summary and conclusion

In this study, psychometric characteristics of seven CANTAB tests were determined. According to the results, the internal consistency was acceptable only in the RTI task. The one-year stability was moderate-to-good for the PRM, SSP, IED, RTI, and RVP. The SRM and SOC tasks were not reliable or stable measures in this study. In addition, the PRM, IED, and RVP had ceiling effects in the present study population of 12-year-old healthy children. The results of this study also suggest that psychometric characteristics of traditional neuropsychological tests may not remain when transplanted into computer form. Thus, the reliability and the stability of CANTAB and other computer-based test batteries among the target population should be confirmed before using them for clinical or research purposes.

References

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*(3), 297–310. doi: 10.1177/107319110000700308

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington DC: American Educational Research Association.

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology, 8*(2), 71–82. doi: 10.1076/chin.8.2.71.8724

Anderson, P., Anderson, V., & Lajoie, G. (1996). The Tower of London test: Validation and standardization for pediatric populatons. *The Clinical Neuropsychologist, 10*(1), 54–65. doi: 10.1080/13854049608406663

Betts, J., Mckay, J., Maruff, P., & Anderson, V. (2006). The development of sustained attention in children: The effect of age and task load. *Child Neuropsychology*, 12(3), 205-221. doi: 10.1080/09297040500488522

Blackwell, A. D., Sahakian, B. J., Vesey, R., Semple, J. M., Robbins, T. W., & Hodges, J. R. (2003). Detecting dementia: novel neuropsychological markers of preclinical Alzheimer's disease. *Dementia and geriatric cognitive disorders*, 17(1-2), 42–48. doi: 10.1159/000074081

Bishop, D., Aamodt-Leeper, G., Creswell, C., McGurk, R., & Skuse, D. (2001). Individual differences in cognitive planning on the Tower of Hanoi task: Neuropsychological maturity or measurement error? *Journal of Child Psychology and Psychiatry, 42*(4), 551–556. doi: 10.1111/1469-7610.00749

Cambridge Cognition Ltd. (2006). CANTABeclipse^{tm}: Software User Guide. Manual version 3. Cambridge: Cambridge Cognition Ltd. (2006).

Cernich, A. N., Brennana, D. M., Barker, L. M., & Bleiberg, J. (2007). Sources of error in computerized neuropsychological assessment. *Archives of Clinical Neuropsychology, 22*(Suppl 1), S39–S48. doi: 10.1016/j.acn.2006.10.004

Conners, C. K., Epstein, J. N., Angold, A., & Klaric, J. (2003). Continuous performance test performance in a normative epidemiological sample. *Journal of Abnormal Child Psychology, 31*(5), 555–562. doi: 10.1023/A:1025457300409

Feldstein, S. N., Keller, F. R., Portman, R. E., Durham, R. L., Klebe, K. J., & Davis, H. P. (1999). A comparison of computerized and standard versions of the Wisconsin card sorting test. *The Clinical Neuropsychologist, 13*(3), 303–313. doi: 10.1076/clin.13.3.303.1744

The Finnish Psychological Test Committee (2013). *Guidelines for using computer-assisted tests in psychological assessment.* [in Finnish]. Available at http://www.psyli.fi/files/1445/Tietokoneavusteinen_psykologinen_tutkimus_18.12.2012.pdf

Fisher, A., Boyle, J., Paton, J., Tomporowski, P., Watson, C., McColl, J., & Reilly, J. (2011). Effects of a physical education intervention on cognitive function in young children: Randomized controlled pilot study. *BMC Pediatrics, 11*(97). doi:10.1186/1471-2431-11-97

Fried, R., Hirshfeld-Becker, D., Petty, C., Batchelder, H., & Biederman, J. (2012). How informative is the CANTAB to assess executive functioning in children with ADHD? A controlled study. *Journal of Attention Disorders*. doi: 10.1177/1087054712457038

Gau, S. S., & Shang, C. (2010). Executive functions as endophenotypes in ADHD: Evidence from the cambridge neuropsychological test battery (CANTAB). *Journal of Child Psychology and Psychiatry, 51*(7), 838–849. doi: 10.1111/j.1469-7610.2010.02215.x

Gnys, J. A., & Willis, W. G. (1991). Validation of executive function tasks with young children. *Developmental Neuropsychology, 7*(4), 487–501. doi: 10.1080/87565649109540507

Halperin, J. M., Sharma, V., Greenblatt, E., & Schwartz, S. T. (1991). Assessment of the continuous performance test: Reliability and validity in a nonreferred sample. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3*(4), 603–608. doi: 10.1037/1040-3590.3.4.603

Henry, L. A., & Bettenay, C. (2010). The assessment of executive functioning in children. *Child and Adolescent Mental Health, 15*(2), 110–119. doi: 10.1111/j.1475-3588.2010.00557.x

Herzog, W., & Boomsma, A. (2009). Small-sample robust estimators of noncentrality-based and incremental model fit. *Structural Equation Modeling*, 16(1), 1-27. doi:10.1080/10705510802561279

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis; Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. doi:10.1080/10705519909540118

Humes, G. E., Welsh, M. C., Retzlaff, P., & Cookson, N. (1997). Towers of Hanoi and London: Reliability of two executive function tasks. *Assessment, 4*(3), 249–257.

Krikorian, R., Bartok, J., & Gay, N. (1994). Tower of London procedure: A standard method and developmental data. *Journal of Clinical and Experimental Neuropsychology, 16*(6), 840–850. doi: 10.1080/01688639408402697

Levaux, M. N., Potvin, S., Sepehry, A. A., Sablier, J., Mendrek, A., & Stip, E. (2007). Computerized assessment of cognition in schizophrenia: promises and pitfalls of CANTAB.

European Psychiatry, 22(2), 104–115. Retrieved from http://dx.doi.org/10.1016/j.eurpsy.2006.11.004

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*(2), 151–173. doi :10.1207/S15328007SEM0902_1

Lowe, C., & Rabbitt, P. (1998). Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: Theoretical and practical issues. *Neuropsychologia, 36*(9), 915–923. doi: 10.1016/S0028-3932(98)00036-0

Luciana, M. (2003). Practitioner review: Computerized assessment of neuropsychological function in children: Clinical and research applications of the cambridge neuropsychological testing automated battery (CANTAB). *Journal of Child Psychology and Psychiatry, 44*(5), 649-663. doi: 10.1111/1469-7610.00152

Luciana, M., Collins, P. F., Olson, E. A., & Schissel, A. M. (2009). Tower of London performance in healthy adolescents: The development of planning skills and associations with self-reported inattention and impulsivity. *Developmental Neuropsychology, 34*(4), 461–475. doi:10.1080/87565640902964540

Luciana, M., Lindeke, L., Georgieff, M., Mills, M., & Nelson, C. A. (1999). Neurobehavioral evidence for working-memory deficits in school-aged children with histories of prematurity. *Developmental Medicine & Child Neurology, 41*(8), 521–533. doi: 10.1111/j.1469-8749.1999.tb00652.x

Luciana, M., & Nelson, C. A. (1998). The functional emergence of prefrontally-guided working memory systems in four-to eight-year-old children. *Neuropsychologia, 36*(3), 273–293. doi: 10.1016/S0028-3932(97)00109-7

Luciana, M., & Nelson, C. A. (2002). Assessment of neuropsychological function through use of the cambridge neuropsychological testing automated battery: Performance in 4-to 12-year-old children. *Developmental Neuropsychology, 22*(3), 595–624. doi :10.1207/S15326942DN2203_3

Mammarella, I. C., Pazzaglia, F., & Cornoldi, C. (2008). Evidence for different components in children's visuospatial working memory. *British Journal of Developmental Psychology, 26*(3), 337–355. doi: 10.1348/026151007X236061

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin; British Medical Bulletin; British Medical Bulletin, 27*(3), 272–277.

Miyake, A., Friedman, N.P., Rettinger, D. A., Shah, P, & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General, 130*(4), 621–540. doi: 10.1037/0096-3445.130.4.621

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory*. (3rd ed.). New York: McGraw-Hill.

Owen, A. M., Downes, J. J., Sahakian, B. J., Polkey, C. E., & Robbins, T. W. (1990). Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia, 28*(10), 1021–1034. Retrieved from http://www.cambridgebrainsciences.com/assets/default/files/pdf/Owen1990.pdf

Parsey, C. M., & Schmitter-Edgecombe, M. (2013). Applications of Technology in Neuropsychological Assessment. *The Clinical neuropsychologist*, (ahead-of-print), 1–34. 10.1080/13854046.2013.834971

Rhodes, S. M., Riby, D. M., Matthews, K., & Coghill, D. R. (2011). Attention-deficit/hyperactivity disorder and Williams syndrome: Shared behavioral and neuropsychological profiles. *Journal of Clinical and Experimental Neuropsychology*, 33(1), 147–156. doi: 10.1080/13803395.2010.495057

Robbins, T. W., James, M., Owen, A. M., Sahakian, B. J., Lawrence, A. D., McInnes, L., & Rabbitt, P. M. (1998). A study of performance on tests from the CANTAB battery sensitive to frontal lobe dysfunction in a large sample of normal volunteers: Implications for theories of executive functioning and cognitive aging. *Journal of the International Neuropsychological Society, 4*(5), 474–490. Retrieved from http://journals.cambridge.org/article_S1355617798455073

Sahakian, B., & Owen, A. (1992). Computerized assessment in neuropsychiatry using CANTAB: Discussion paper. *Journal of the Royal Society of Medicine, 85*(7), 399–402. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1293547/pdf/jrsocmed00109-0031.pdf

Shallice, T., & Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London.B, Biological Sciences, 298*(1089), 199–209. doi: 10.1098/rstb.1982.0082

Smith, P. J., Need, A. C., Cirulli, E. T., Chiba-Falek, O., & Attix, D. K. (2013). A comparison of the Cambridge automated neuropsychological test battery (CANTAB) with "traditional" neuropsychological testing instruments. *Journal of Clinical and Experimental Neuropsychology,* (ahead-of-print), 1–10. doi:10.1080/13803395.2013.771618

Westland, C.J. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*, 9(6), 476-487. doi:10.1016/j.elerap.2010.07.003

Table 1

CANTAB tests measuring different dimensions of cognitive functions.

Dimension of cognitive function	Test	Abbreviation	Description	Score
Visual memory	1. Pattern Recognition Memory	PMR	Children had to remember the presented geometric patterns and discriminate them from the novel patterns.	No. of correct responses (max 24)
	2. Spatial Recognition Memory	SRM	Children had to remember the location of white squares and discriminate these locations from the novel locations.	No. of correct responses (max 20)
Executive function	3. Spatial Span	SSP	Specified number of white boxes changed their color one by one and children had to reproduce the same sequence by touching the boxes in the same order the boxes changed their color.	Span length (max 9)
	4. Stockings of Cambridge	SOC	Children had to move the colored balls in the lower part of the screen to the same position in the stockings as they were in the upper part of the screen. They had specified number of moves to use.	No. of problems solved in minimum moves (max 12)
	5. Intra-Extra Dimensional Set Shift	IED	The children had to choose one of the two different dimensions: one is correct, and the other is incorrect. According to immediate feedback, they had to choose the correct pattern and learn the rule.	No. of stages completed
Attention	6. Reaction Time	RTI	Children had to hold down the press pad button until they saw the yellow spot flashing on one of the five circles, and then touch the middle of the circle as quickly as possible.	5-choice reaction time (ms), 5-choice movement time (ms)
	7. Rapid Visual Information Processing	RVP	Children had to touch the press pad button every time they discriminate the target sequence (digits 3, 5, and 7) from the digits appearing in a pseudo-random order at the rate of 100 digits per minute.	A' (range: .00 to 1.00; bad to good) = how well child detect the target sequences

Table 2

Mean values, standard deviations (SD), and gender differences for CANTAB tests in the first assessment in 2011

Measurements in 2011	Boys (*n*=99)			Girls (*n*=131)			All (*n*=230)			*p* [a]
	Mean	SD	%	Mean	SD	%	Mean	SD	%	
PRM no. of correct responses (max 24)	20.88	2.80		20.84	2.29		20.86	2.52		.478
SRM no. of correct responses (max 20)	16.49	1.61		16.81	1.67		16.67	1.65		.152
SSP span length (max 9)	6.53	1.27		6.68	1.37		6.61	1.33		.385
SOC no. of problems solved in minimum moves (max 12)	7.69	1.81		7.61	1.71		7.64	1.75		.746
RTI five-choice movement time (ms)	329	74		364	87		349	83		.001
RTI five-choice reaction time (ms)	300	36		318	32		310	35		<.001
RVP A' (range 0–1)	0.97	0.02		0.97	0.03		0.97	0.02		.973
IED % of children who completed at least stage 8			76			73			74	.154

Note. Abbreviations: PRM, Pattern Recognition Memory; SRM, Spatial Recognition Memory; SSP, Spatial Span; SOC, Stockings of Cambridge; RTI, Reaction Time; RVP, Rapid Visual Information Processing; IED, Intra-Extra Dimensional Set Shift.

[a] *P*-value for gender differences.

Table 3

Mean values and standard deviations (SD) for all variables for children who participated in measurements in 2011 and 2012 and effects sizes (Cohen's d) for the repeated measures

Measurements	Boys (*n*=27)			Girls (*n*=47)			All (*n*=74)		
2011	Mean	SD	*d*	Mean	SD	*d*	Mean	SD	*d*
PRM no. of correct responses (max 24)	20.96	3.39	.22	20.68	2.30	.10	20.78	2.73	.14
SRM no. of correct responses (max20)	16.63	1.45	.02	16.70	1.79	.14	16.68	1.66	.10
SSP span length (max 9)	6.52	1.19	.32	6.28	1.35	.55	6.36	1.29	.46
SOC no. of problems solved in minimum moves (max 12)	7.52	1.76	.43	7.38	1.69	.60	7.43	1.71	.54
RTI five-choice movement time (ms)	321	53	.05	370	83	.16	352	76	.12
RTI five-choice reaction time (ms)	304	43	-.15	319	30	.05	314	36	-.02
RVP A' (range 0–1)	.97	.02	.35	.96	.03	.40	.96	.03	.38
IED % of children who completed at least stage 8	75.8			73.3			74.3		
2012	Boys (*n*=27)			Girls (*n*=47)			All (*n*=74)		
PRM no. of correct responses (max 24)	21.56	1.65		20.96	2.80		21.18	2.45	
SRM no. of correct responses (max20)	16.67	1.52		16.94	1.51		16.84	1.51	
SSP span length (max 9)	6.93	1.07		6.98	1.00		6.96	1.00	
SOC no. of problems solved in minimum moves (max 12)	8.26	1.72		8.40	1.39		8.35	1.51	
RTI five-choice movement time (ms)	325	65		382	75		361	76	
RTI five-choice reaction time (ms)	298	34		321	36		313	37	
RVP A' (range 0–1)	.98	.02		.98	.03		.98	.03	
IED % of children who completed at least stage 8	88.2			75.9			80.7		

Note. Abbreviations: PRM, Pattern Recognition Memory; SRM, Spatial Recognition Memory; SSP, Spatial Span; SOC, Stockings of Cambridge; RTI, Reaction Time; RVP, Rapid Visual Information Processing; IED, Intra-Extra Dimensional Set Shift; *d*, Cohen's d.

Table 4

Proportion (%) of children in 2011 and 2012 with different Spatial Span (SSP) lengths (2–9)

Measurement year	SSP span length						
	3	4	5	6	7	8	9
In 2011	1.3	1.3	22.6	20.4	23.9	24.8	5.7
In 2012	0	0	11.4	17.0	39.8	27.3	4.5

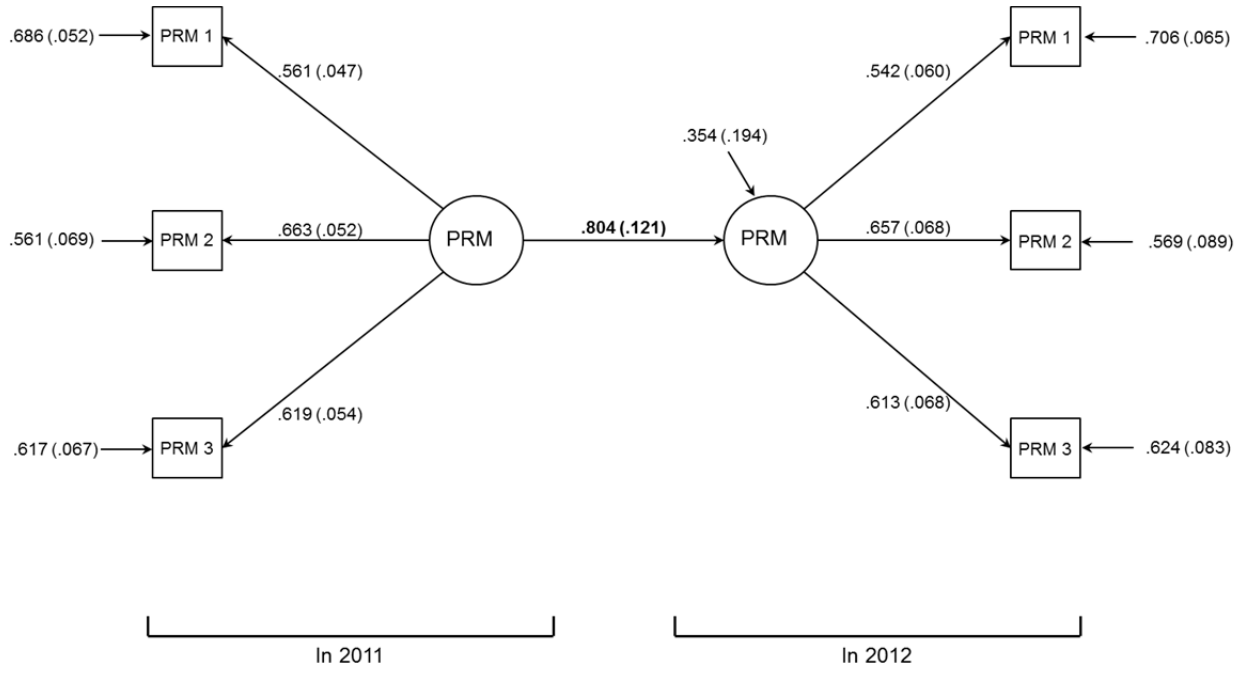

Figure 1. The estimation results of the stability model of the Pattern Recognition Memory (PRM) test for 2011 and 2012. Standardized parameter estimates and standard errors are presented. Three parcels from 24 patterns (incorrect/correct response) of the PRM test were constructed and used as the indicator variables in the confirmatory factor analyses (PRM 1, PRM 2, PRM 3). All the factor loadings were constrained to be equal to one.

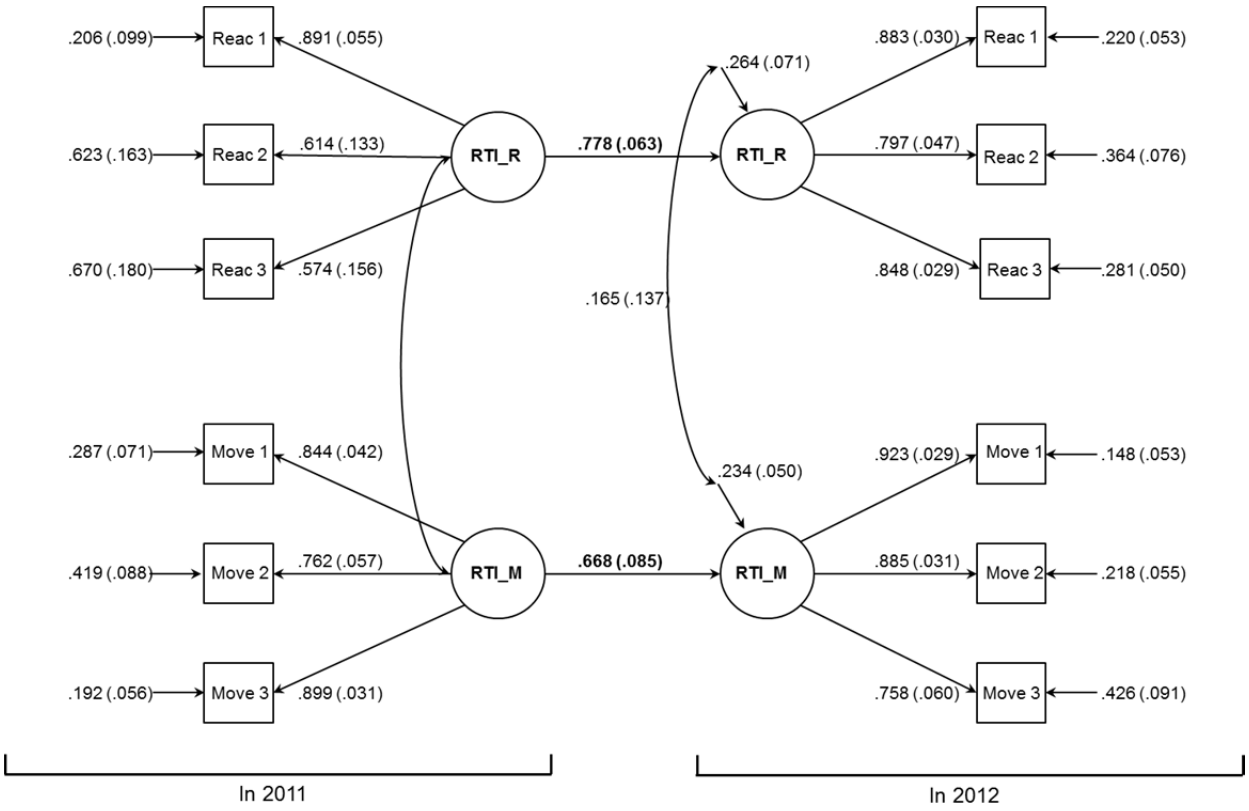

Figure 2. The stability model for the Reaction Time (RTI) test for 2011 and 2012. Standardized parameter estimates and standard errors are presented. For structural equation modeling, three subscales of the reaction time (Reac 1, Reac 2, Reac 3) and movement time (Move 1, Move 2, Move 3) at both measurement points were formed from 15 patterns of RTI. All the factor loadings were constrained to be equal to one.

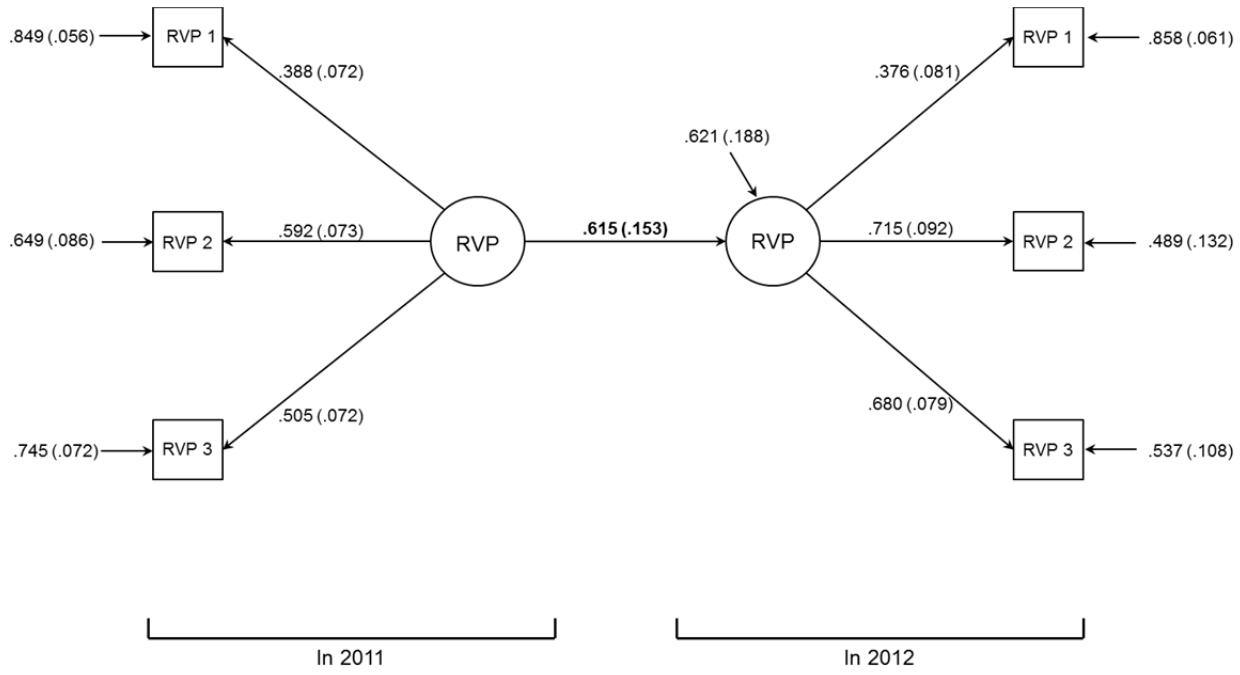

Figure 3. The stability model for the Rapid Visual Information Processing (RVP) test for 2011 and 2012. Standardized parameter estimates and standard errors are presented. The RVP (multiplied by 10) of the three blocks of the test (RVP 1, RVP 2, RVP 3) were used as indicator variables in the structural equation modeling. Factor loadings were constrained to be equal across the time points.