Behav Res (2015) 47:708–719
DOI 10.3758/s13428-014-0469-8

Validity and reliability of an online visual–spatial working memory task for self-reliant administration in school-aged children

Eva Van de Weijer-Bergsma · Evelyn H. Kroesbergen · Emilie J. Prast · Johannes E. H. Van Luit

Published online: 26 April 2014

Abstract Working memory is an important predictor of academic performance, and of math performance in particular. Most working memory tasks depend on one-to-one administration by a testing assistant, which makes the use of such tasks in large-scale studies time-consuming and costly. Therefore, an online, self-reliant visual–spatial working memory task (the Lion game) was developed for primary school children (6–12 years of age). In two studies, the validity and reliability of the Lion game were investigated. The results from Study 1 (n = 442) indicated satisfactory six-week test–retest reliability, excellent internal consistency, and good concurrent and predictive validity. The results from Study 2 (n = 5,059) confirmed the results on the internal consistency and predictive validity of the Lion game. In addition, multilevel analysis revealed that classroom membership influenced Lion game scores. We concluded that the Lion game is a valid and reliable instrument for the online computerized and self-reliant measurement of visual–spatial working memory (i.e., updating).

Keywords Visual-spatial · Working memory · Computerized classroom assessment · Children

Working memory (WM) is the capacity to store and manipulate information for brief periods of time and is an important predictor for academic performance in areas such as reading (De Weerdt, Desoete & Roeyers 2013; Gathercole, Alloway, Willis & Adams 2006; Swanson, Xinhua & Jerman 2009) and mathematics in particular (Bull, Espy & Wiebe 2008; Friso-van den Bos, Van der Ven, Kroesbergen & Van Luit 2013; Swanson, 2006; Toll, Van der Ven, Kroesbergen & Van Luit 2011; Van der Ven, Kroesbergen, Boom & Leseman 2012). The most widely used model of WM includes several components (Baddeley, 2000; Baddeley & Hitch, 1974). Two slave systems, the visuospatial sketchpad and the phonological loop, involve temporary storage of visual and spatial information, and phonological and auditory information, respectively. The slave systems are usually measured with simple span tasks, in which increasingly longer strings of information must be reproduced. More recently, a third slave system was added to the model: the episodic buffer. The episodic buffer is a temporary storage system that is responsible for the integration of information from a variety of sources (Baddeley, 2000). The central executive coordinates information within these slave systems, and is traditionally measured with complex span tasks. In these tasks, storage as well as processing or manipulation of information is required. In other words, WM can be distinguished from short-term memory, which only involves the temporary storage of information by the slave systems, whereas WM involves storage as well as processing of information. More recently, the coordinating role of the central executive has been differentiated further into different subprocesses. On the basis of Baddeley's model and the executive function literature, the subprocesses of inhibition, shifting and updating are distinguished (Miyake et al. 2000). *Inhibition* is the ability to suppress a dominant response in favor of another response or no response at all. Shifting is the ability to switch between response sets. *Updating* is the ability to monitor and adjust the information that is active in working memory. Factor analysis has confirmed the distinction between these—interrelated—executive functions (Miyake et al. 2000). The recent review by Friso-van den Bos et al. indicated that, although the three components of working memory are related to math performance, updating seems most strongly related with mathematical performance of all executive functions.

E. Van de Weijer-Bergsma (✉) · E. H. Kroesbergen · E. J. Prast · J. E. H. Van Luit
Faculty of Social and Behavioral Sciences, Department of Pedagogical and Educational Sciences, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands
e-mail: e.vandeweijer@uu.nl

Individual differences in WM capacity have important consequences for learning (Alloway & Alloway, 2010). Poor WM leads to failures in remembering and keeping up with lesson instructions, choosing the right strategy for a math problem and keeping track of the sequence of steps in a complex strategy (Alloway, 2006). On the other hand, children who are better able to hold relevant information in mind and manipulate this information, have an advantage when it comes to using prior knowledge and procedures in mathematics (Van der Ven, Boom, Kroesbergen & Leseman 2012; Van der Ven, Kroesbergen, Boom and Leseman 2012b). WM can be assessed with a wide variety of measures, including behavioral rating scales filled in by teachers or parents (e.g., the Behavior Rating Inventory of Executive Function [BRIEF; Gioia, Isquith, Guy & Kenworthy 2000] or the Working Memory Rating Scale [WMRS; Alloway, Gathercole, Kirkwood & Elliott 2009]), paper-and-pencil tests (e.g., Working Memory Test Battery for Children [WMTB-C]; Pickering & Gathercole, 2001), and computerized tests (e.g., Automated Working Memory Assessment [AWMA]; Alloway, Gathercole, Kirkwood & Elliott 2008). Behavioral rating scales give information on WM functioning in daily life settings, such as family life or classroom. Also, they are generally rapidly administered and easily scored, and therefore not very time-consuming. Standardized performance-based tests (i.e., paper-and-pencil or computerized) give a more objective representation of differences between individuals than behavioral ratings, since the former are not influenced by the subjective experience of the rater. Also, although overall behavioral ratings are associated with performance-based tests, ratings for more specific executive functions often do not correlate with their performance-based counterpart (Alloway et al. 2009; Mahone et al. 2002; Mangeot, Armstrong, Colvin, Yeates & Taylor 2002; Toplak, Bucciarelli, Jain & Tannock 2008; Vriezen & Pigott, 2002). However, one-to-one administration is very time-consuming and costly. In research with large samples, WM assessment is often unfeasible due to money and time constraints. Also, since standardized performance-based tests are usually individually administered in a quiet room by a (trained) testing assistant, the testing situation differs greatly from the real-life classroom situation in which children learn academic skills. In the classroom, WM capacity can be influenced by multiple factors such as ambient noise (Baker & Holding, 1993; Stansfeld et al. 2005), classroom distractors (e.g., classmates walking around) and emotional interference (e.g., general anxiety, perfectionism, performance, or test anxiety; Ashcraft & Kirk, 2001; Dutke & Stöber, 2001; Hadwin, Brogan & Stevenson 2005). Therefore, we have developed a WM task (Lion game) that children can start, run, and finish autonomously within the classroom setting, without the presence of a testing assistant. The task can also be administered in groups. The feasibility of computerized or online, self-reliant, and therefore group-administrable WM tasks has been shown in adult samples (De Neys, d'Ydewalle, Schaeken & Vos 2002; Pardo-Vázquez & Fernández-Rey, 2008). Computerized operation span tasks have been found to be a valid and reliable measure of WM in Flemish (De Neys et al. 2002), Spanish (Pardo-Vázquez & Fernández-Rey, 2008), and American samples (Redick et al. 2012). The reliability and validity of two computerized, visual–spatial WM tasks has also been examined in a university sample as well as in 11- to 14-year-old children (McPherson & Burns, 2008). Although the concurrent and predictive validity was found to be good for one of these tasks, results for the second task were not acceptable in the school-aged sample. Recently, the predictive validity of an online adaptive computerized visual–spatial working memory task for math abilities was demonstrated in a large sample of primary school children, from Grades 1 to 6 (Van der Ven, Van der Maas, Straatemeier & Jansen 2013). In the present studies, we investigated the validity and reliability of the Lion game, a visual–spatial updating task developed for primary school children from Grades 1 to 6. First, a detailed description of the development and characteristics of the Lion game will be given. Second, a study investigating the concurrent validity, predictive validity and test–retest reliability of the Lion game will be described (Study 1). Third, a study will be described in which the predictive validity of the Lion game was investigated in a very large sample (Study 2). Since classroom factors, such as ambient noise, may influence task performance, we also examined classroom effects in this large sample.

Development and characteristics of the Lion game

The Lion game is an online computerized visual–spatial complex span task, in which children have to search for colored lions. The task is adapted from a WM training for children in kindergarten and Grade 1 (Kolkman, Hoijtink, Kroesbergen & Leseman 2013; Kolkman, Kroesbergen & Leseman 2011). Children are presented with a 4 × 4 matrix containing 16 bushes (see Fig. 1). In each trial, eight lions of different colors (red, blue, green, yellow, and purple) are consecutively presented at different locations in the matrix for 2,000 ms. Children have to remember the *last* location where a lion of a certain color has appeared, and use the mouse button to click on that location after the sequence has ended. The task consists of five levels, in which working memory load is manipulated by the number of colors—and hence, the number of locations—that children have to remember and update. At Level 1, children have to remember the location of the last red lion. At Level 2, children have to remember the locations of the last red and the last blue lion, and so on (Level 3: red, blue, and yellow; Level 4: red, blue, yellow, and green; Level 5: red, blue, yellow, green, and purple). Items were constructed using randomization with regard to sequence of location and color, with one constraint: items never end with a red lion, since the first response requires the location of the last red lion. Before starting the task, all children are presented with

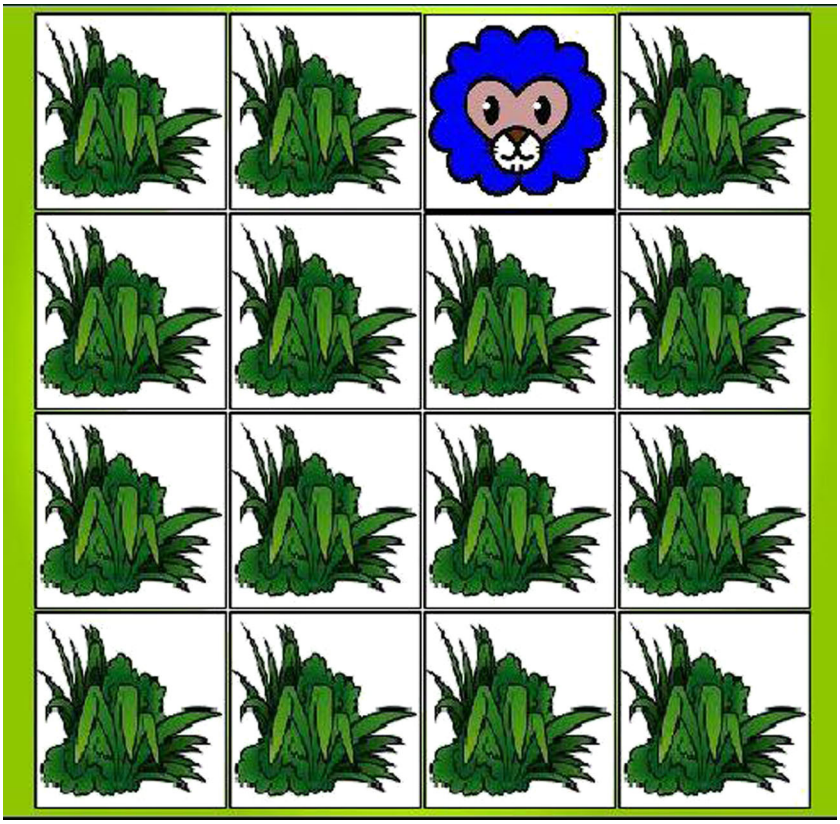

Fig. 1 Snapshot of the 4 × 4 matrix with bushes of the Lion game

two practice trials, in which they have to remember the locations of the last red and last blue lion. After each practice trial, children receive feedback on their performance. When children fail both practice trials, a third practice trial is presented. Item scores are the sum of the numbers of correct responses within an item. Because the number of lions to be remembered increases with each level, the maximum score also increases with each level, ranging from a maximum score of 1 (at Level 1) to 5 (at Level 5). The proportion of items recalled in the correct serial position of the task is scored (St. Clair-Thompson & Sykes 2010). Since the task requires children to remember the last location of an increasing number of differently colored lions, the Lion game taps into updating skills. However, as in most complex span tasks, the Lion game probably also taps into set-shifting (e.g., because of the different colors used) and inhibition (e.g., because children have to inhibit the previous location where a target lion appeared).

The task was tested in a small pilot study to examine feasibility and to inspect responses for item selection for the final version of the task. A total of 75 children from Grades 1–6 from four different elementary schools in the Netherlands participated. During the pilot, the different levels consisted of six trials each. Children were presented with only three of the six levels. That is, n = 38 children in the first and second grade finished Levels 1, 2, and 3; n = 29 children in the third and fourth grades finished Levels 2, 3, and 4; and n = 8 children in the fifth and sixth grades finished Levels 3, 4, and 5. Analysis of the results from the pilot revealed that items within the same level showed similar response patterns. Mean scores and distribution of scores were inspected and from each level, four out of six items were selected for inclusion in the final version of the task. Although all items within a level showed similar patterns, the two items that diverged the most from same-level items or showed the most overlap with adjacent-level items were excluded. From each level, four trials were selected for the revised version of the task. In the final version of the task, all children are presented with all five levels, consisting of four items each.

Study 1

The goal of this study was to examine the test–retest reliability, internal consistency, concurrent validity, and predictive validity of the Lion game.

Method

Participants

A total of 442 children from Grades 1–6 from two elementary schools from the eastern part of The Netherlands participated. From each class, a subsample of eight children was selected

for the administration of tester-led WM tasks ($n = 98$), administered to examine concurrent validity. Systematic sampling with an interval of six children from an alphabetically ordered student list per class was used for this selection process. At the end of the list, counting was continued from the top of the list, skipping children who had already been selected. Table 1 presents the sample characteristics. Parents of children received written information on the study and were notified about the voluntary nature of participation. A passive consent procedure was used: parents informed the head of the school of their child when they did not want their child to participate. All children received passive consent. The study was approved by the ethics committee of the Faculty of Social and Behavioral Science, Utrecht University.

Measurements

Lion game See the description of the task in the "Development and Characteristics of the Lion Game" section.

Automated Working Memory Test (AWMA) A Dutch version of the AMWA (Alloway et al. 2008) was used. Two subtests were used to measure visual–spatial WM: dot matrix and odd-one-out. Both tasks consist of blocks of six items, with blocks increasing in difficulty. The dot matrix consists of nine blocks, the odd-one-out of seven blocks. After four correct items, the next block is presented. After three incorrect answers within a block, task administration is ended. In the dot matrix task, children have to recall the position of one or more red dots in a series of four by four matrices. The number of dots presented increases with each successive block. Children indicate in which squares the dots appeared by tapping the squares on the screen. Children fail an item when one or more of the positions is not recalled correctly or when the positions are tapped in the wrong order. This task requires children to store visual–spatial and temporal information about the appearing dots in short-term memory, and is therefore traditionally considered a simple span task tapping into the visuospatial sketchpad component in Baddeley's model of working memory. However, the dynamic presentation of the Dot matrix also requires visual tracking and controlled processing by the central executive (Alloway, Gathercole & Pickering 2006; Miyake, Friedman, Rettinger, Shah & Hegarty 2001). In the odd-one-out task, children view three shapes, each encased in a square presented in a row. Children must first determine the odd-one-out shape. At the end of the trial, children must recall the location of all odd one out shapes, in the correct order, by tapping the correct squares on the screen. The number of shape sets is increased with each successive block. The odd-one-out task is a complex span task that requires children to identify and remember the location of odd shapes, while keeping information about previous trials in memory, and is therefore considered to tap into the visuospatial sketchpad as well as the central executive component of Baddeley's working memory model. Scores on both tasks are calculated on the basis of the number of finished blocks (six points per block) and the number of correct answers within the last unfinished block. Maximum scores for the dot matrix and odd-one-out are 54 and 42, respectively. The AWMA subtests are shown to be reliable and valid measures of WM (Alloway & Alloway, 2010; Alloway et al. 2008; Injoque-Ricle, Calero, Alloway & Burin 2011). In the original version, a score of one standard deviation below the mean is interpreted as indicating problems in working memory (Alloway et al. 2008).

Math performance Mathematical abilities were measured using the criterion-based Cito Mathematics Tests (CMT; Janssen, Scheltens & Kraemer 2005a). These are national Dutch tests with good psychometric properties that are commonly used in Dutch schools to monitor the progress of primary school children. Two different versions were created for each grade, one to be administered at mid school year (M) and one at the end of the school year (E), except for Grade 6, which has a test at the beginning of the school year (B6) and one at mid school year (M6). From M1 (i.e., mid-Grade 1) through M6 (i.e., mid-Grade 6), five main domains are covered: (a) numbers and number relations, covering the structure of the number line and relations between numbers, (b) addition and subtraction, (c) multiplication and division, (d) complex math applications, often involving multiple mathematical manipulations, and (e) measuring (e.g., weight and length). From M2 to M6, several domains are added successively: (f) estimation, (g) time, (h) money, (i) proportions, (j) fractions, and (k) percentages. Raw scores are converted into competence scores that increase throughout primary school, enabling the comparison of the results of different versions (Janssen, Scheltens & Kraemer 2005b). The reliability coefficients of the different

Table 1 Sample characteristics of Studies 1 and 2

	Study 1		Study 2
	Total Sample (*n*)	Subsample* (*n*)	(*n*)
Total	442	98	5,059
Grade 1	61	15	847
Grade 2	80	15	824
Grade 3	68	16	825
Grade 4	58	15	846
Grade 5	85	18	838
Grade 6	90	19	879
% of girls	50.5	51.0	49.3
	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)
Age (years)	9.7 (1.8)	9.5 (1.7)	9.2 (1.8)

* Subsample of children selected with a six-interval systematic sampling procedure

versions range from .91 to .97 (Janssen, Verhelst, Engelen & Scheltens 2010).

Procedure

Administration of working memory tasks occurred within a period of four weeks in February 2013. The Lion game was group-administered in the classroom. The teacher gave a general instruction:

> You will be doing a lion game on the computer. Everyone in the class has their own pictogram with their name written under it. You can play the game by clicking on your own name or pictogram. Use the headset to hear the instruction. Finish the game without pausing. If you need to use the bathroom, please do so before you start.

A research assistant was present during administration to check proper functioning of the game. The odd-one-out and dot matrix tasks were administered individually by one of two research assistants in a quiet room in the school building. Math performance tests were administered as part of the regular school testing procedure at mid school year and results were requested from the teacher. In March 2013, six weeks after the first administration, the Lion game was administered for the second time in the subsample (see Table 1 for the sample characteristics) according to the same procedure.

Data analysis

No univariate or multivariate outliers were identified using z scores (criterion: z score > 3.29) and Mahalanobis distances [$\chi^2(2) = 13.816$], respectively. Normality of distributions of the variables was examined by calculating the standardized skewness and kurtosis index (statistic divided by standard error). Values were found to be higher than 3 for the Lion game and CMT scores, indicating that the distributions differed significantly from normality. Therefore, Spearman's Rho correlations (one-tailed) with related constructs were calculated to examine concurrent and predictive validity. Spearman's Rho correlations between different measurement occasions were calculated to examine test–retest reliability.

The internal consistency of the scores was calculated in two different ways following the methods used by Engle and coworkers (Engle, Tuholski, Laughlin & Conway 1999; Kane et al. 2004). First, because there are four trials within each level, each trial can be identified as the first, second, third, or fourth instance of a particular difficulty. We calculated a sum score for the first trials within the five levels, and calculated the same for the second, third and fourth trials within the five levels. Then Cronbach's alpha was calculated between these scores (Engle et al. 1999). In the second approach, Cronbach's alpha was calculated for the proportion correct scores on each individual trial (Kane et al. 2004).

Since the analysis included categorical variables as control variables, univariate analyses of covariance (ANCOVAs) were used to investigate the predictive value of Lion game mean proportion correct scores (independent variable) for math performance on the CMT (dependent variable), after controlling for grade and classroom effects. Classroom is a categorical variable and was included as a fixed factor. Since the increase in mean score on the dependent variable CMT score with increasing grade was not linear, Grade was also included as a fixed factor.

To explore age-related differences in the validity of the task, correlational analyses and ANCOVAs were also run for children in Grades 1 through 3 (i.e., younger children) and children in Grades 4 through 6 (i.e., older children) separately.

Missing values

Of the 442 children who initially completed the Lion game, 68 children missed the second assessment six weeks later, leaving n = 374 children for the analysis of test–retest reliability. Although some children missed the second assessment due to absence from school, the majority of the children missed the second assessment due to technical problems with the website.

Results

Test–retest reliabilities for the Lion game were $\rho = .70$ ($p < .001$) for the mean proportion scores and $\rho = .71$ ($p < .001$) for the absolute scores. In addition, although proportions correct as well as absolute scores were significantly higher at Time 2 than at Time 1 for each task (all ps < .01), the mean increase in absolute scores was only one to two points, indicating relatively small practice effects on the Lion game. Importantly, as is indicated by the high test–retest reliabilities, the rank-ordering of individuals was stable across test sessions.

Cronbach's alpha was calculated for the Lion game as an index of internal consistency. Internal consistency for the sum scores of each first, second, third, and fourth instance within the different levels was α= .86. The proportion correct scores for each individual item revealed an internal consistency of α = .87.

Performance on the Lion game was significantly related to performance on the AWMA odd-one-out ($\rho = .51$, $p < .001$) and the AWMA dot matrix ($\rho = .59$, $p < .001$) tasks. Correlations between the Lion game and the two AWMA tasks were similar to the correlations among the two AWMA tasks ($r = .56$, $p < .001$). Exploration of age-related differences revealed that correlations were stronger in younger children ($\rho = .57$, $p < .001$, and $\rho = .55$, $p < .001$, respectively) than in older children ($\rho = .25$, $p < .05$, and $\rho = .28$, $p < .05$,

respectively). The same pattern of results was found, however, for the relationship between the two AWMA tasks, with a stronger relationship in younger children (r = .52, p < .001) than in older children (r = .26, p < .05).

The ANCOVA results showed that performance on the Lion game was a significant predictor of math performance on the CMT, $F(1, 422) = 51.53$, $p < .001$, $\eta_p^2 = .11$, after controlling for grade and classroom effects. Exploring age-related differences revealed that the predictive value of Lion game performance for math performance was larger in younger children, $F(1, 199) = 30.35$, $p < .001$, $\eta_p^2 = .13$, than in older children, $F(1, 222) = 18.43$, $p < .001$, $\eta_p^2 = .07$.

Study 2

The goals of this study were (1) to investigate the effects of classroom membership on Lion game performance and (2)to replicate the results regarding predictive validity of the Lion game in a large sample, using multilevel analysis.

Method

Participants

This study was part of a large-scale intervention study on the effects of teacher training in differentiated math education on student math performance. A total of 5,237 children from Grades 1 through 6 from 32 elementary schools in The Netherlands participated. Parents of all children received written information about the study and a passive informed consent procedure was used. Parents informed the teacher of their child or a designated contact person at their school when they did not want their child to participate. Thirteen children did not receive parental consent to participate. The study was approved by the ethics committee of the Faculty of Social and Behavioral Science, Utrecht University.

Measures

The same measures were used to assess working memory (i.e., the Lion game) and math performance (i.e., Cito Mathematics Test) that had been used in Study 1.

Procedure

At the beginning of 2012–2013 school year (September–October 2012), as part of a pretest for the intervention study, teachers received an automated e-mail containing login information for their class of children and were asked to let all of the students within their class perform the Lion game within a period of three weeks. Math performance tests (CMT) were administered as part of the regular school testing procedure, and results were requested from the mid-school-year results (January–February 2013).

Power analysis

Power and sample size were determined with a Monte Carlo study using the Mplus program (Muthén & Muthén, 2002). In a Monte Carlo study, data are generated from a population with hypothesized parameter values. A large number of samples are drawn (replications), and for each sample a model is estimated, with parameter values and standard errors averaged over samples (Muthén & Muthén, 2002). In the present Monte Carlo study, model estimation was carried out using a nonnormality-robust sandwich estimator, calculating standard errors to deal with nonnormal data. A two-level model with one dependent variable (i.e., math performance) and two independent variables (i.e., grade and working memory) was selected for the design. The results from Study 1 were used to estimate parameter values. Using 1,000 replications, the criteria of (1) less than 10 % parameter and standard error bias, (2) less than 5 % standard error bias for the parameter for which power was being assessed, and (3) coverage between .91 and .98 were met, resulting in 100 % power, with a required overall sample size of 1,875 students and 75 classes.

Available data

From the 5,237 children participating, 5,059 children provided data for the analyses. Table 1 presents the sample characteristics. The Lion game was finished by 4,588 children, and for 4,462 of these children we received mid-school-year CMT scores at the time of analysis. A total of 3,986 children from 216 classes had data on both variables. On the basis of the power analysis, we can conclude that this sample was large enough for the model specified above. Children who were included were compared to children who were excluded from the analyses, and they did not differ with regard to gender, age, or grade.

Data analysis

No univariate or multivariate outliers were identified using z scores (criterion: z score > 3.29) and Mahalanobis distances [$\chi^2(2) = 13.816$], respectively. The normality of the variable distributions was examined by calculating standardized skewness and kurtosis indexes (statistic divided by standard error). These values were found to be higher than 3 for both the Lion game and CMT scores, indicating that the distributions differed significantly from normality.

The internal consistency of the scores was calculated for the whole sample as well as for each grade separately in two different ways, following the methods used by Engle and

coworkers (Engle et al. 1999; Kane et al. 2004). More details can be found in Study 1.

Since the data had a nested structure, with children being nested in classrooms, multilevel analysis with Mplus (Muthén & Muthén, 2006) was used. The number of schools in this study (32 schools) was too small to add school as a separate level of the analysis (Hox, Van de Schoot & Matthijsse 2012), and therefore a two-level structure was used (Level 1, individual children; Level 2, class). In all models, we controlled for grade. A full estimation maximum likelihood (MLR) method was used, since it is robust to nonnormality and can handle missing data.

First, in a two-level multilevel model, the intraclass correlation (ICC) was calculated for Lion game mean proportion correct scores to indicate the ratio of variance *between* classes to variance *within* those classes, using grade as a control variable. Following Hox (2002), ICC values of .05, .10, and .15 are considered to be small, medium, and large, respectively (Hox, 2002). However, besides the size of the ICC value, a design effect greater than 2 would indicate that the clustering in the data needed to be taken into account during estimation. A design effect was calculated by 1 + (average cluster size – 1) * intraclass correlation (Muthén & Satorra, 1995). Second, CMT was added to the model, also using grade as a control variable, but still without CMT being regressed on Lion game score. Third, CMT score was regressed on Lion game score with a fixed slope, to investigate the predictive validity of the Lion game for math performance. Fourth, the slope was allowed to be random, to investigate whether the relationship between Lion game and CMT scores varied between classes.

By convention, models have a good fit if χ^2 is low, RMSEA is less than .05, and CFI and TLI are close to 1 (Arbuckle, 2006). Because of the large sample size, however, we expected the χ^2 test to be significant. When two nested models are compared, the difference between their respective –2LL values is evaluated using a likelihood ratio test.

Results

Results for the Lion game and CMT scores are presented in Table 2.

Cronbach's alpha was calculated for the Lion game as an index of internal consistency. Internal consistency for the sum scores of each first, second, third, and fourth instance within the different levels was α = .89 for the whole sample and ranged from .82 to .86 in the different grades separately. The proportion correct scores for each individual item revealed an internal consistency of α = .90 for the whole sample and ranged from .80 to .85 in the different grades separately.

First, an ICC of .07 was found for Lion game proportion correct scores after controlling for grade, which indicated that a small proportion of variance was explained by class membership. The design effect for Lion game scores [1 + (21.94 – 1) * .07 = 2.46] was greater than 2, indicating that clustering in the data needed to be taken into account during estimation. Children in higher grades received significantly higher scores on the Lion game (standardized estimate [SE] = .57, $p < .001$) and the CMT (SE = .78, $p < .001$).

Second, when adding CMT score to the model, an ICC of .05 was found for CMT scores. In this model, 79 % of the variance in CMT scores was explained by grade and classroom membership.

Third, Lion game performance was found to be a significant predictor of CMT scores (SE = .19, $p < .001$), after controlling for grade.

Fourth, when the regression slope of CMT score on Lion game score was allowed to vary, a slope mean of 0.3 ($p < .001$) and a slope variance of 0.007 ($p < .001$) was found (95 % confidence interval: .004–.10). This indicates that the strength of the relationship between visual–spatial working memory and math ability varied significantly between classes. However, since the slope variance was small, we compared the model with the random slope (–2LL = 11,430) to the model in which the slope was fixed (–2LL = 11,445). The model with a fixed slope provided a better fit [Δ{–2LL(1)} = 58, $p < .001$]. This final model fit the data well [$\chi^2(1) = 12.85$, $p < .001$, RMSEA = .048, CFI = .974, TLI = .896] and explained 82 % of the variance in CMT scores. The proportion of variance explained by Lion game score was 3 %.

Finally, we explored age-related differences in the predictive value of Lion game performance for math performance by running the final model in each grade separately. In most grades, this model showed a good fit, except for Grades 3 and 4 (see Table 3 for standardized estimates, ICCs, and model fit indices). Most importantly, the results showed that Lion game performance explained 16 %, 13 %, 13 %, 11 %, 8 %, and 5 % of the variance in math performance in Grades 1, 2, 3, 4, 5, and 6, respectively.

Discussion

The aim of this research was to investigate the validity and reliability of an online visual–spatial complex span task (the Lion game) for self-reliant administration in school-aged children. Taken together, the results from two studies showed that the Lion game has good internal consistency reliability, satisfactory test–retest reliability, and good concurrent and predictive validity. In addition, classroom membership influenced working memory performance.

With regard to test–retest reliability, the results from the Lion game are comparable to the working memory tasks from the AWMA, for which scores vary between .64 and .80 (Alloway et al. 2006). These results are promising, especially when we consider the fact that tester-led tasks such as the

Table 2 Means and standard deviations (*SD*) for outcome measures in Study 1 and Study 2

	Study 1								Study 2				
	Lion Game			CMT		AWMA			Lion Game			CMT	
	n	Proportion Correct Score	Absolute Score	*n*	Ability Score	*n*	Dot Matrix	Odd-One-Out	*n*	Proportion Correct Score	Absolute Score	*n*	Ability Score
Grade 1	61	.52 (.18)	25.49 (10.72)	61	35.00 (14.41)	15	19.80 (4.39)	13.27 (5.88)	744	.46 (.18)	22.89 (9.49)	738	34.77 (16.08)
Grade 2	80	.61 (.15)	30.80 (8.28)	80	53.35 (13.42)	15	20.80 (2.73)	15.53 (4.70)	757	.56 (.18)	28.64 (10.04)	750	53.23 (15.38)
Grade 3	68	.70 (.11)	35.48 (7.61)	68	73.00 (12.44)	16	23.44 (4.78)	17.94 (4.04)	751	.65 (.17)	33.76 (10.13)	703	73.67 (15.17)
Grade 4	58	.75 (.11)	39.93 (8.13)	58	88.05 (11.22)	15	24.33 (3.18)	19.67 (4.76)	744	.71 (.15)	37.31 (9.38)	742	86.90 (13.72)
Grade 5	85	.77 (.09)	40.30 (7.17)	85	100.56 (11.55)	18	28.28 (4.25)	20.78 (3.94)	787	.74 (.13)	39.34 (8.93)	743	100.83 (12.03)
Grade 6	90	.78 (.10)	41.63 (7.15)	90	112.89 (9.23)	19	29.05 (3.76)	22.10 (4.76)	802	.77 (.13)	41.47 (8.79)	786	110.23 (14.77)
Total	442	.70 (.16)	36.02 (9.89)	442	79.60 (29.50)	98	24.59 (5.20)	18.45 (5.50)	4,588	.56 (.19)	34.03 (11.43)	4,462	76.96 (30.16)

CMT = Cito Math Test, AWMA = Automated Working Memory Assessment Battery

AWMA are administered under controlled circumstances (one-to-one administration in a quiet room), whereas the Lion game was administered in groups in a classroom setting that may vary in for example classroom order and atmosphere, disturbances or ambient noise between different assessments.

The concurrent validity of the Lion game with other visual–spatial working memory tasks was good, which indicated that the task requires storage as well as processing. The relationship between performance on all three tasks (i.e., Lion game, dot matrix, and odd-one-out) was stronger in younger children (Grades 1–3) than in older children (Grades 4–6). The most apparent explanation for these results is that variation in task performance was somewhat smaller in older children. However, another—not mutually exclusive—explanation might be that the processes that these tasks tap into change with development. It has been argued that even simple span tasks require more controlled processing in children than in older children or adults (Alloway et al. 2006; Jarvis & Gathercole, 2003), due to less automated rehearsal and chunking in younger children (Engle et al. 1999). In addition, there are indications that the different components of executive functions (i.e., updating, set-shifting and inhibition) are less separable and/or more strongly interrelated in younger children (Van der Ven, Kroesbergen, et al., 2012; Wiebe, Espy & Charak 2008). So, it might be that different working memory tasks and the demands that they make on (sub)processes are more similar in younger children than in older children.

Theoretically, the Lion game has a clear updating component. However, like other updating tasks, it also requires inhibition. In *N*-back tasks, for example, such as the letter memory task used by Miyake et al. (2001), participants have to remember the last four letters of a list. This task requires participants to add the most recent letter and drop (and inhibit) the 5th letter back. The Lion game is probably more complex than such *N*-back tasks, since children also have to use different colors as categories, which may demand some set-shifting abilities, as well. As such, the Lion game is more comparable to the "keep track" task, in which participants are asked to remember the last words in several categories (e.g., animals, colors, countries; Miyake et al. 2001).

An advantage of the Lion game may be its higher ecological validity, since the task is administered in the same classroom environment in which learning takes place. The same

Table 3 Standardized estimates of multilevel models examining math on Lion game, intraclass correlations (ICC) and fit indices for each grade separately

	Standardized Estimate	ICC Lion Game	ICC Math	χ^2	*df*	*p*	CFI	TLI	RMSEA
Grade 1	.40*	.08	.14	132.26	2	.000	.963	.925	.08
Grade 2	.56*	.07	.12	38.29	2	.000	.998	.996	.01
Grade 3	.36*	.10	.18	98.28	2	.000	.669	.338	.20
Grade 4	.33*	.10	.13	197.15	2	.000	.877	.754	.17
Grade 5	.28*	.09	.08	54.21	2	.000	.955	.910	.05
Grade 6	.22*	.07	.31	32.82	2	.000	.990	.980	.02

* $p < .01$, CFI = comparative fit index, TLI = Tucker–Lewis index, RMSEA = root mean square error of approximation

factors that may influence learning, such as classroom disturbances, for example, may influence performance on the Lion game. However, this may also be considered a potential confound and may influence the validity of what is being measured by the Lion game. Other differences between children in, for example, distractibility, attention, motivation, or the ability to work independently may affect task performance more in the classroom setting than in one-to-one administration by a testing assistant or teacher.

Consistent with the literature, we found a significant age effect on visual–spatial working memory (Alloway & Alloway, 2010; Alloway et al. 2006; Van der Ven, Kroesbergen, et al., 2012). Children from higher grades received higher scores on the Lion game.

Regarding the predictive value of the Lion game, we found visual–spatial working memory scores to be significantly predictive of later math achievement, which is consistent with the previous literature (Raghubar, Barnes & Hecht 2010; Van der Ven et al. 2013). Although we did not control for intelligence scores, previous research has shown that working memory is predictive of academic performance above and beyond intelligence scores (De Weerdt et al. 2013; Swanson & Beebe-Frankenberger, 2004). In addition, our results revealed that the predictive value of the Lion game for math performance declined with age. This finding is consistent with results from previous studies, which indicate that the predictive value of visual–spatial working memory for math achievement changes with age (Friso-van den Bos et al. 2013; Imbo & Vandierendonck, 2007; McKenzie, Bull & Gray 2003; Raghubar et al. 2010; Van der Ven et al. 2013). It is often suggested that these results reflect the fact that younger children who learn and apply new mathematical skills rely more on visual–spatial working memory, whereas older children increasingly rely on verbal working memory after skills have been learned. In our study, however, we cannot exclude the possibility that this result was an artifact of smaller variation in Lion game performance and math performance in older children.

The finding that performance on the Lion game was affected by classroom membership indicates that, indeed, an effect of classroom variables does need to be taken into account. Which variables affect classroom differences is still unclear, and more research will be needed to identify those factors at different levels. At the school level, for example, factors such as the location of the school (e.g., ambient noise as a result of a location near to a road) and setting in a low socioeconomic neighborhood may account for differences between classes. At the class level, social climate (e.g., focus on achievement) and teacher variables (e.g., classroom management, need supporting) may be important. Such environmental factors are particularly interesting, since working memory ability is not fixed, and its development can be influenced by environmental factors, such as parenting (Bernier, Carlson & Whipple 2010; Dilworth-Bart, Poehlmann, Hilgendorf, Miller & Lambert 2010). The effects of working memory training programs are less clear, although short-term, specific training effects have been found (Melby-Lervåg & Hulme, 2013). At the student level, student characteristics (e.g., more or fewer students with trait anxiety, or low socioeconomic status) may influence classroom differences. Although we also found significant classroom effects on the *strength* of the relationship between working memory and subsequent math performance, the effect was small, and a model in which this variation was fixed provided a better fit. This indicates that this finding was most likely due to the large power of the study, and its practical relevance is probably minimal.

The Lion game is a low-cost measure, enabling the inclusion of working memory as a control or predictor variable in large sample studies. Also, the game can be easily translated into other languages, making cross-cultural comparisons possible without difficulty. The results in the data from the Lion game per age group from Study 2 are included in Table 4 for use by other researchers. A link to the Dutch version of the online task can be requested from the first author. It must be stated that the possibilities for interpreting individual results and using the Lion game for diagnostic purposes are limited. However, the game might perhaps be used as a quick screening instrument for working memory problems. The sensitivity of the task to clinical indications should be investigated further. Clinical sensitivity might be increased by developing a larger battery of tasks. We are currently developing a verbal recall backward task that can be administered online in the same age groups. The reliability and validity for this task will be investigated, as well.

Some of the task features of the Lion game need further discussion. First, the game-like structure makes the task more attractive to children than are many currently used tasks, making it more user-friendly. However, children who are more experienced in computer-based work or play could have an advantage over children who are less experienced, which

Table 4 Means and standard deviations (*SD*s) for the Lion game per age group

Age	*n*	Proportion Correct Score [Mean (*SD*)]	Absolute Score [Mean (*SD*)]
5 years	49	.442 (.190)	22.1 (9.7)
6 years	621	.471 (.181)	23.4 (9.6)
7 years	722	.557 (.186)	28.4 (10.4)
8 years	703	.649 (.171)	33.7 (10.1)
9 years	720	.697 (.155)	36.6 (9.6)
10 years	755	.734 (.145)	38.9 (9.4)
11 years	771	.754 (.142)	40.3 (9.4)
12 years	193	.738 (.157)	39.7 (9.9)
13 years	4	.638 (.254)	35.3 (12.0)

may influence task performance. Second, no cutoff rules were used in the Lion game. Although this can make the task more frustrating for children who have difficulties working self-reliantly or who have working memory difficulties, it also increases the sensitivity of the task to individual differences. In contrast, when testing is ended, once accuracy falls below a certain threshold or cutoff rule, information on all following trials is discarded, and the sensitivity of a task becomes limited (Conway, Kane, Bunting, Hambrick, Wilhelm and Engle 2005). Third, although the verbal instructions as part of the task were kept as straightforward as possible, it is possible that children with fewer verbal skills had more difficulty understanding the task. It is currently unknown whether such child characteristics (e.g., computer experience or verbal abilities) confound task performance. However, in Study 1, research assistants were present during testing. They reported that, although most children successfully and self-reliantly finished the task, a few younger children closed the Web browser unintentionally, having to start over. Also, the children who were selected as a subsample in Study 1 were asked about their experience with the AWMA tasks and the Lion game. None of these children reported difficulties with understanding the instruction on any of the tasks.

To conclude, this study showed that it is possible to use an online self-reliant computer program to reliably and validly measure visual–spatial working memory in a classroom setting. Because this method allowed for data collection in a large sample, it was possible to show that classroom membership influenced task performance.

Author note This study is financed by The Netherlands Organisation for Scientific Research (NWO), Grant No. 411-10-753. We gratefully acknowledge the contribution by Tanja van Veldhuizen, Elien van Es, Manon Kroeze en Ilse Egberdink in the (organization of) data collection. We thank Levent Serbesatik for programming the Lion game for Web-based administration. Mirjam Moerbeek is gratefully acknowledged for her assistance in conducting power analysis.

References

Alloway, T. P. (2006). How does working memory work in the classroom? *Educational Research and Reviews, 1,* 134–139.

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology, 106,* 20–29. doi:10.1016/j.jecp.2009.11.003

Alloway, T. P., Gathercole, S. E., Kirkwood, H., & Elliott, J. (2008). Evaluating the validity of the Automated Working Memory Assessment. *Educational Psychology, 28,* 725–734. doi:10.1080/01443410802243828

Alloway, T. P., Gathercole, S. E., Kirkwood, H., & Elliott, J. (2009). The working memory rating scale: A classroom-based behavioral assessment of working memory. *Learning and Individual Differences, 19,* 242–245. doi:10.1016/j.lindif.2008.10.003

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? *Child Development, 77,* 1698–1716. doi:10.1111/j.1467-8624.2006.00968.x

Arbuckle, J. L. (2006). *Amos 7.0 user's guide*. Chicago: Amos Development Corp.

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety and performance. *Journal of Experimental Psychology: General, 130,* 224–237. doi:10.1037/0096-3445.130.2.224

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4,* 417–423. doi:10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York: Academic Press.

Baker, M. A., & Holding, D. H. (1993). The effects of noise and speech on cognitive task performance. *Journal of General Psychology, 120,* 339–355. doi:10.1080/00221309.1993.9711152

Bernier, A., Carlson, S. M., & Whipple, N. (2010). From external regulation to self-regulation: Early parenting precursors of young children's executive functioning. *Child Development, 81,* 326–339. doi:10.1111/j.1467-8624.2009.01397.x

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology, 33,* 205–228. doi:10.1080/87565640801982312

Conway, A. A., Kane, M. J., Bunting, M., Hambrick, D. Z., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12,* 769–786. doi:10.3758/BF03196772

De Neys, W., d'Ydewalle, G., Schaeken, W., & Vos, G. (2002). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica, 42,* 177–190.

De Weerdt, F., Desoete, A., & Roeyers, H. (2013). Working memory in children with reading disabilities and/or mathematical disabilities. *Journal of Learning Disabilities, 46,* 461–472. doi:10.1177/0022219412455238

Dilworth-Bart, J., Poehlmann, J., Hilgendorf, A. E., Miller, K., & Lambert, H. (2010). Maternal scaffolding and preterm toddlers' visual–spatial processing and emerging working memory. *Journal of Pediatric Psychology, 35,* 209–220. doi:10.1093/jpepsy/jsp048

Dutke, S., & Stöber, J. (2001). Test anxiety, working memory, and cognitive performance: Supportive effects of sequential demands. *Cognition and Emotion, 15,* 381–389. doi:10.1080/02699930125922

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128,* 309–331. doi:10.1037/0096-3445.128.3.309

Friso-van den Bos, I., Van der Ven, S. H. G., Kroesbergen, E. H., & Van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review, 10,* 29–44. doi:10.1016/j.edurev.2013.05.003

Gathercole, S. E., Alloway, T. P., Willis, C., & Adams, A. (2006). Working memory in children with reading disabilities. *Journal of Experimental Child Psychology, 93,* 265–281. doi:10.1016/j.jecp.2005.08.003

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). Test review: Behavior Rating Inventory of Executive Function. *Child Neuropsychology, 6,* 235–238. doi:10.1076/chin.6.3.235.3152

Hadwin, J. A., Brogan, J., & Stevenson, J. (2005). State anxiety and working memory in children: A test of processing efficiency theory. *Educational Psychology, 25,* 379–393. doi:10.1080/01443410500041607

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah: Erlbaum.

Hox, J., Van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods, 6,* 87–93.

Imbo, I., & Vandierendonck, A. (2007). The development of strategy use in elementary school children: Working memory and individual differences. *Journal of Experimental Child Psychology, 96,* 284–309. doi:10.1016/j.jecp.2006.09.001

Injoque-Ricle, I., Calero, A. D., Alloway, T. P., & Burin, D. I. (2011). Assessing working memory in Spanish-speaking children: Automated Working Memory Assessment battery adaptation. *Learning and Individual Differences, 21,* 78–84. doi:10.1016/j.lindif.2010.09.012

Janssen, J., Scheltens, F., & Kraemer, J. M. (2005a). *Rekenen-wiskunde groep 3–8: Handleidingen [Mathematics test first through sixth grade manuals]*. Arnhem: Cito.

Janssen, J., Scheltens, F., & Kraemer, J. M. (2005b). *Leerling- en onderwijsvolgsysteem rekenen-wiskunde [Student monitoring system mathematics]*. Arnhem: Cito.

Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS rekenen-wiskunde voor groep 3 tot en met 8 [Scientific justification of the mathematics test for grade 1 through grade 6]*. Arnhem: Cito.

Jarvis, H. L., & Gathercole, S. E. (2003). Verbal and non-verbal working memory and achievements on National Curriculum tests at 11 and 14 years of age. *Educational and Child Psychology, 20,* 123–140.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133,* 189–217. doi:10.1037/0096-3445.133.2.189

Kolkman, M. E., Hoijtink, H. J. A., Kroesbergen, E. H., & Leseman, P. P. M. (2013). The role of executive functions in numerical magnitude skills. *Learning and Individual Differences, 24,* 145–151. doi:10.1016/j.lindif.2013.01.004

Kolkman, M. E., Kroesbergen, E. H., & Leseman, P. P. M. (2011). *The impact of verbal and visual working memory training on numerical skills*. Paper presented at the symposium of the Society for Research in Child Development, Montreal.

Mahone, E. M., Cirino, P. T., Cutting, L. E., Cerrone, P. M., Hagelthorn, K. M., Hiemenz, J. R., & Denckla, M. B. (2002). Validity of the behavior rating inventory of executive function in children with ADHD and/or Tourette syndrome. *Archives of Clinical Neuropsychology, 17,* 643–662. doi:10.1016/S0887-6177(01)00168-8

Mangeot, S., Armstrong, K., Colvin, A. N., Yeates, K. O., & Taylor, H. G. (2002). Long-term executive function deficits in children with traumatic brain injuries: Assessment using the Behavior Rating Inventory of Executive Function (BRIEF). *Child Neuropsychology, 8,* 271–284. doi:10.1076/chin.8.4.271.13503

McKenzie, B., Bull, R., & Gray, C. (2003). The effects of phonological and visual–spatial interference on children's arithmetical performance. *Educational and Child Psychology, 20,* 93–108.

McPherson, J., & Burns, N. (2008). Assessing the validity of computer-game-like tests of processing speed and working memory. *Behavior Research Methods, 40,* 969–981. doi:10.3758/BRM.40.4.969

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49,* 270–291. doi:10.1037/a0028228

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41,* 49–100. doi:10.1006/cogp.1999.0734

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,* 621–640. doi:10.1037/0096-3445.130.4.621

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9,* 599–620. doi:10.1207/S15328007SEM0904_8

Muthén, L. K., & Muthén, B. O. (2006). *Mplus*. Los Angeles: Muthén & Muthén.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25,* 267–316.

Pardo-Vázquez, J., & Fernández-Rey, J. (2008). External validation of the computerized, group administrable adaptation of the "operation span task.". *Behavior Research Methods, 40,* 46–54. doi:10.3758/BRM.40.1.46

Pickering, S. J., & Gathercole, S. E. (2001). *Working memory test battery for children*. London: Psychological Corp.

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences, 20,* 110–122. doi:10.1016/j.lindif.2009.10.005

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment, 28,* 164–171. doi:10.1027/1015-5759/a000123

St. Clair-Thompson, H. L., & Sykes, S. (2010). Scoring methods and the predictive ability of working memory tasks. *Behavior Research Methods, 42,* 969–975. doi:10.3758/BRM.42.4.969

Stansfeld, S., Berglund, B., Clark, C., Lopez-Barrio, I., Fischer, P., Öhrström, E., & Berry, B. (2005). Aircraft and road traffic noise and children's cognition and health: A cross-national study. *Lancet, 365,* 1942–1949. doi:10.1016/S0140-6736(05)66660-3

Swanson, H. L. (2006). Cross-sectional and incremental changes in working memory and mathematical problem solving. *Journal of Educational Psychology, 98,* 265–281. doi:10.1037/0022-0663.98.2.265

Swanson, H. L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology, 96,* 471–491. doi:10.1037/0022-0663.96.3.471

Swanson, H. L., Xinhua, Z., & Jerman, O. (2009). Working memory, short-term memory, and reading disabilities: A selective meta-analysis of the literature. *Journal of Learning Disabilities, 42,* 260–287. doi:10.1177/0022219409331958

Toll, S. W. M., Van der Ven, S. H. G., Kroesbergen, E. H., & Van Luit, J. E. H. (2011). Executive functions as predictors of math learning disabilities. *Journal of Learning Disabilities, 44,* 521–532. doi:10.1177/0022219410387302

Toplak, M. E., Bucciarelli, S. M., Jain, U., & Tannock, R. (2008). Executive functions: Performance-based measures and the Behavior Rating Inventory of Executive Function (BRIEF) in adolescents with attention deficit/hyperactivity disorder (ADHD). *Child Neuropsychology, 15,* 53–72. doi:10.1080/09297040802070929

Van der Ven, S. H. G., Boom, J., Kroesbergen, E. H., & Leseman, P. P. M. (2012a). Microgenetic patterns of children's multiplication learning: Confirming the overlapping waves model by latent growth modeling. *Journal of Experimental Child Psychology, 113,* 1–19. doi:10.1016/j.jecp.2012.02.001

Van der Ven, S. H. G., Kroesbergen, E. H., Boom, J., & Leseman, P. P. M. (2012b). The development of executive functions and early mathematics: A dynamic relationship. *British Journal of Educational Psychology, 82,* 100–119. doi:10.1111/j.2044-8279.2011.02035.x

Van der Ven, S. H. G., Van der Maas, H. L. J., Straatemeier, M., & Jansen, B. R. J. (2013). Visuospatial working memory and mathematical

ability at different ages throughout primary school. *Learning and Individual Differences, 27,* 182–192. doi:10.1016/j.lindif.2013.09.003

Vriezen, E. R., & Pigott, S. E. (2002). The relationship between parental report on the BRIEF and performance-based measures of executive function in children with moderate to severe traumatic brain injury. *Child Neuropsychology, 8,* 296–303. doi:10.1076/chin.8.4.296.13505

Wiebe, S. A., Espy, K. A., & Charak, D. (2008). Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. *Developmental Psychology, 44,* 575–587. doi:10.1037/0012-1649.44.2.575