

This article was downloaded by: [Karolinska Institute]

On: 28 September 2010

Access details: Access Details: [subscription number 779857390]

Publisher Psychology Press

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Developmental Neuropsychology

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t775653638>

### Measuring Working Memory Capacity With Greater Precision in the Lower Capacity Ranges

Sissela Bergman Nutley<sup>a</sup>; Stina Söderqvist<sup>a</sup>; Sara Bryde<sup>a</sup>; Keith Humphreys<sup>b</sup>; Torkel Klingberg<sup>a</sup>

<sup>a</sup> Neuropaediatric Research Unit, Department of Women's and Children's Health, Stockholm Brain Institute, Karolinska Institutet, Stockholm, Sweden <sup>b</sup> Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Online publication date: 17 December 2009

**To cite this Article** Nutley, Sissela Bergman , Söderqvist, Stina , Bryde, Sara , Humphreys, Keith and Klingberg, Torkel(2010) 'Measuring Working Memory Capacity With Greater Precision in the Lower Capacity Ranges', *Developmental Neuropsychology*, 35: 1, 81 – 95

**To link to this Article: DOI:** 10.1080/87565640903325741

**URL:** <http://dx.doi.org/10.1080/87565640903325741>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

## Measuring Working Memory Capacity With Greater Precision in the Lower Capacity Ranges

Sissela Bergman Nutley, Stina Söderqvist, and Sara Bryde  
*Neuropaediatric Research Unit, Department of Women's and Children's Health,  
and Stockholm Brain Institute, Karolinska Institutet, Stockholm, Sweden*

Keith Humphreys  
*Department of Medical Epidemiology and Biostatistics, Karolinska Institutet,  
Stockholm, Sweden*

Torkel Klingberg  
*Neuropaediatric Research Unit, Department of Women's and Children's Health,  
and Stockholm Brain Institute, Karolinska Institutet, Stockholm, Sweden*

Working memory capacity is usually measured as the number of stimuli correctly remembered. However, these measures lack precision when assessing individuals with low capacity. This study aimed to create a more precise measure of visuospatial working memory capacity, using intra-level differences in difficulty between items. In two experiments, children aged 4–6 years ( $N = 97$ ) were tested on a large number of items. Data showed a large variability of difficulty within each level and the factors contributing to this variability were identified. This variability can be used to provide a precise measure of working memory capacity in the lower ranges.

Visual-spatial working memory (VSWM) is the ability to remember visually presented spatial information during a short period of time. One way of measuring VSWM capacity is with spatial-span tasks, in which dots or blocks are marked in a sequential order to be repeated by the subject in the same way. Spatial span normally varies from two to three objects for 4-year-olds and three to five objects for 6-year-olds (Gathercole, Pickering, Ambridge, Wearing, 2004; Nichelli, Bulgheroni, & Riva, 2001). In young children, performance on VSWM tasks can predict future development of abilities such as reasoning, math, and reading (Gathercole, Brown, & Pickering 2003; Bull, Espy, & Wiebe, 2008; Alloway, Gathercole, Willis, Adams, 2004). Several neurodevelopmental disorders, including attention deficit hyperactivity disorder (ADHD) are associated with deficits in VSWM (Westerberg, Hirvikoski, Forssberg, Klingberg, 2004). There is therefore a great need for sensitive measures of WM capacity, especially in young chil-

---

This study was supported by Knut and Alice Wallenbergs stiftelse and Riksbankens Jubileumsfond.

Correspondence should be addressed to Sissela Bergman Nutley, MR-Centrum, Karolinska Sjukhuset, N8, 171 76 Stockholm, Sweden. E-mail: sissela.bergman@ki.se

dren in which this measure can be used for identifying children at risk for future academic difficulties.

One of the most commonly used spatial span tasks is the Corsi Block tapping task (Milner, 1971). The Corsi Block tapping task consists of nine blocks semi-randomly placed on a board and the subject has to repeat visuospatial sequences by tapping blocks in the correct order from memory. Span tasks have been argued to be preferred when studying developmental differences since it is possible to use the same task across different capacities (Gathercole et al., 2004). However, in these tasks, difficulty is always increased by adding one additional object to be remembered at each level (defined as items consisting of the same number of objects). This is problematic at lower spans, as the proportionate increase of remembering three instead of two objects is 50%, whereas a person with a capacity of remembering six objects only receives a proportionate increase of 17% in order to remember seven objects. This illustrates a discrepancy in the precision across different levels of this type of test and specifically highlights a reduction in the lower capacity ranges. One solution to this problem is to identify sub-levels with differences in difficulty that are finer than the levels defined by the number of objects. Previous research has shown that in the Corsi Block tapping task (or adapted versions of it), the path configuration of a sequence affects the difficulty (Busch, Farrell, Lisdahl-Medina, Krikorian, 2005; Kemps, 2001; Orsini, Pasquadibisceglie, & Picone, 2001; Orsini, Simonetta, & Marmorato, 2004; Parmentier, Elford, & Maybery, 2005; Smirni, Villardita, & Zappalá, 1983). Specific factors found contributing to the difficulty of an item are the number of internal crossings of a pattern (Busch et al., 2005, Orsini et al., 2001), and the distance between objects (Orsini et al., 2004). Although factors that affect difficulty have been identified, little attempt has been made to use this information in a constructive fashion. The overall aim of this study was to use this information in order to create a VSWM test for more precise measurement in the lower capacity ranges.

The goal of this study was therefore to verify the existence of factors that contribute to difficulty, other than the number of objects, and then to use this information in order to create sub-levels that provide a more sensitive measure of VSWM. We expected to replicate previously reported factors contributing to item difficulty (number of internal crossings of a pattern and distance between objects) and to find additional ones. To our knowledge, these factors have not been studied in a child population. The Corsi block tapping task has been adapted and computerized in different versions. We used a common variation of this, which we will refer to as the grid task (GT), in which a  $4 \times 4$  grid appears where dots are presented in a sequential order to be repeated by the subject in the same order (Alloway, Gathercole, & Pickering, 2006; Westerberg et al., 2004).

In Experiment 1 children aged 4–6 years were tested with the GT on a large number of items within each level (consisting of the same number of stimuli to remember) in order to investigate whether there were differences in difficulty within a level as tested with a Rasch model. Difficulties of the items were estimated and the factors contributing to these difficulties were identified with logistic regression. Based on this information, a new test was created, consisting of a subset of items from Experiment 1. In Experiment 2 a new sample of 4–6-year-olds were tested on these items to confirm the progressive difficulty of the sub-levels as tested with a Rasch model. In order to validate the test, the results were compared to another VSWM test; the Odd One Out (OOO) test from the Automated Working Memory Assessment battery (Alloway, 2007).

## EXPERIMENT 1

### Method

#### Subjects

The subjects were recruited from preschools in the Stockholm area. Informed consent was obtained from all of the subjects' parents. A total of 22 four-year-olds with a mean age of 53 months ( $SD = 3.9$ , range 46–59 months, 12 girls) and 15 six-year-olds with a mean age of 76 months ( $SD = 3.2$ , range 69–81 months, 9 girls) participated in the study. All of the children were Caucasian and none of the children had any neurological or psychiatric diagnosis as determined by teacher reports.

#### Task Description

The GT task was based on a previously used VSWM task (Alloway et al., 2006; Westerberg et al., 2004), and programmed in Eprime 1.1 (Psychological Software Tools, Inc). Dots appear in sequential order for the subject to repeat by pointing at the screen (Figure 1). The task was presented on a HP Compaq nc6320 laptop with a 15-inch screen. The dots were yellow and were presented for 1,000 msec, with an interstimulus interval (ISI) of 500 msec. The items were selected from a database of results from children who had undergone WM training on a similar task (Klingberg, Forssberg, & Westerberg, 2002; Klingberg et al., 2005; Thorell, Lindqvist, Bergman Nutley, Bohlin, Klingberg, 2009; database from Cogmed Systems Inc.). Based on previous results we selected stimulus configurations of seven items with success rates in the lowest 10%, seven items in the highest 10%, and seven items around the median from each level. In total 21 items per level were selected for testing, assumed to represent three different sub-levels of difficulty.

#### Procedure

All the subjects were tested at their preschool in a separate and quiet room with the computer monitor placed approximately 50 cm from the child at an appropriate angle. Since the children were too young to handle a mouse, the responses were made by pointing at the screen and performance was manually checked by the test leader on to a separate scoring sheet. The

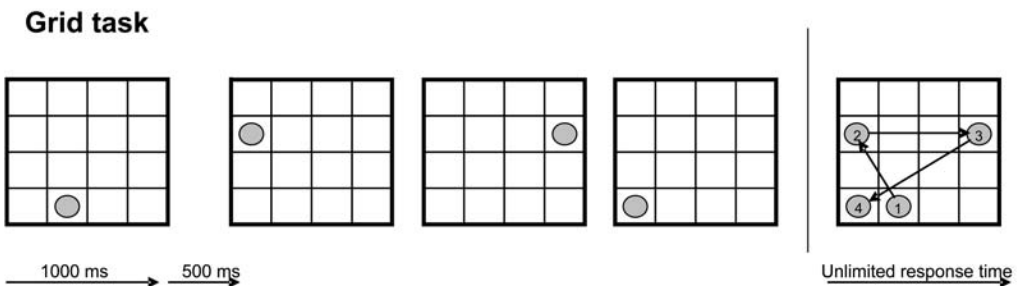


FIGURE 1 The Grid task illustrating an example of a pattern held in memory where (1), (2) and (3) are a side dots, (4) is a corner dot, (2) and (3) appear on the same axis consecutively, whereas (1) and (4) do not, and the connection of dots (3) and (4) causes one crossing of the pattern. The mean distance of the pattern is the mean length of the arrows.

4-year-olds were tested on the GT as part of a larger test battery and were tested on 4–6 different occasions taking breaks whenever the test leader saw fit, each session lasting between 20 and 40 min. The 6-year-olds were tested on 3–5 different occasions based on the same principal with sessions lasting between 30 and 60 min. The younger children started on the level of remembering sequences consisting of two dots and then proceeded to successively higher levels. The test proceeded until the subjects had failed three subsequent items and then completed the items on the sub-level they were currently on and one additional sub-level (consisting of seven items). The 6-year-old subjects started on level three, since it was presumed that they would reach ceiling scores on level two. For all failed and all subsequent items after the termination rule, a zero score was given to these items in order to calculate success rates for each item.

### *Analysis*

Results on the items were analyzed in terms of success rate. The average number of correct responses on each item served as a measurement of success rate (i.e., the number of children who passed an item divided by the number of children who attempted it).

*Rasch analysis.* The data was analyzed using a one parameter item response model, the Rasch model. The Rasch model assumes that the probability of a person passing a particular test item is determined by two parameters: the ability of the person ( $\theta_n$ ) and the difficulty of the test item ( $\delta_i$ ). Not all items were performed by all children. In order to account for this in assessing item difficulties, and to obtain ability scores for each person, further statistical analysis was conducted. The data was analysed using a one parameter item response model, the Rasch model (Equation 1 in the Appendix) for dichotomous data using WINSTEPS (version 3.63.0; ). A value of the reliability in terms of Cronbachs  $\alpha$  is also determined when conducting the Rasch analysis. The parameters were set to generate an item measurement on the logit scale in which the mean item was set to a measurement of 50 and higher scores indicated higher difficulty. For an overview on the Rasch measurement model, see Bond and Fox (2001) and for in depth description see Appendix 1.

*Logistic regression.* Each item was characterized in terms of its constituting factors: the number of dots in a item, the mean number of corner positions of an item, the mean number of corner positions in an item, the mean distance between dots in an item, the mean number of times the pattern line crossed itself, the mean number of times a position was repeated, and the mean number of times a dot was followed by another dot on the same vertical or horizontal axis as the previous one (Figure 1). In order to get an estimate of each factor's contribution to item difficulty independent of the number of dots in the item, the factor was divided by the number of dots in the item.

Forward multiple logistic regression analysis was conducted in SPSS (SPSS for Windows, Rel. 16.0.1. 2007. Chicago: SPSS Inc). All of the factors described earlier were entered into a logistic regression analysis as explanatory variables. The dependent variable was the subject's response on each item; pass or fail. We added a constant to the null model and entered a person factor as an explanatory variable in to the analysis to control for the individual differences that were not of interest in identifying the factors contributing to difficulty (e.g., age, motivation, alertness). The goal was to identify the factors contributing to the probability of passing a specific item.

Results

Figure 2A shows the success rates for the 4-year-olds and the 6-year-olds, respectively, and Figure 2 illustrates the large variability within levels and the overlaps between levels.

The success rate analysis revealed a large variance between items consisting of the same number of dots and showed an overlap between levels consisting of different number of dots, which illustrates that there is more determining the difficulty of an item than just the number of dots.

The Rasch analysis was conducted on the full set of raw data (84 items) and item measurements showed satisfactory fit to the data. Six of the most difficult items showed no calibration value because of maximum estimated measurement explained by the fact that these items had no correct responses and thus there is no further information available separating the difficulties of these items. None of the other items showed misfit according to fit statistics described in Appendix 1 (all mean square were between 0.6 and 1.4 with an associated Z-value of  $\approx 2$ ). Therefore, the test seemed to measure one unidimensional construct. The person separation reliability was estimated to 0.93 and the item separation reliability to 0.89, indicating satisfactory differentiation ability of the test. Cronbach's alpha was estimated from the person's raw scores to 0.94. Four participants showed misfit in their estimated ability scores (two with less variance and two with more variance than expected in their responses).

The logistic regression analysis provided significant evidence that the following factors contributed to the item difficulties: number of dots, mean distance between dots in an item, mean number of side positions of an item, mean number of times the pattern line crossed itself, and mean number of times a dot was followed by another dot on the same axis as the previous one. None of these vari-

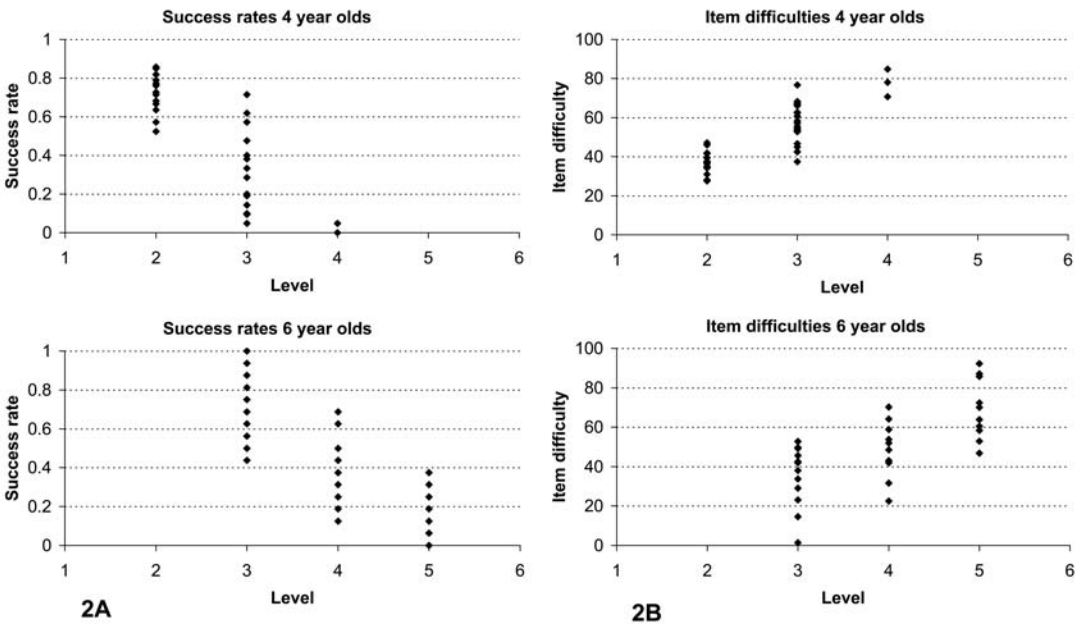


FIGURE 2 (A) Empirical data showing success rates of trials on levels 2-4 as performed by the 4-year-old subjects (top) and on levels 3-5 as performed by the 6-year-old subjects (bottom). (B) Estimates of scaled item difficulties ( $\delta_i$  in Equation 1 in Appendix 1) from a Rasch analysis.

ables could be removed without significantly reducing the fit of the model and no additional variables explained more variance. Model selection was conducted according to goodness of fit statistics (stepwise selected nested model compared to the final model;  $\chi^2$  diff.; 285.6,  $df = 41$ ,  $p < .001$ ).

Having examined which factors contribute to difficulty our next aim was to use this information to create sub-levels that provide a more sensitive measure of VSWM than one only taking levels into account. We therefore selected items that constituted sub-levels of difficulty according to the Rasch analysis, and at the same time had a low error value in the logistic regression, that is, where difference difficulty could be explained by the identified factors (Figure 3A). In other words, we have not only chosen items that differ in difficulty but we also know why they differ. We selected items so that the probabilities of passing were not overlapping between sub-levels ac-

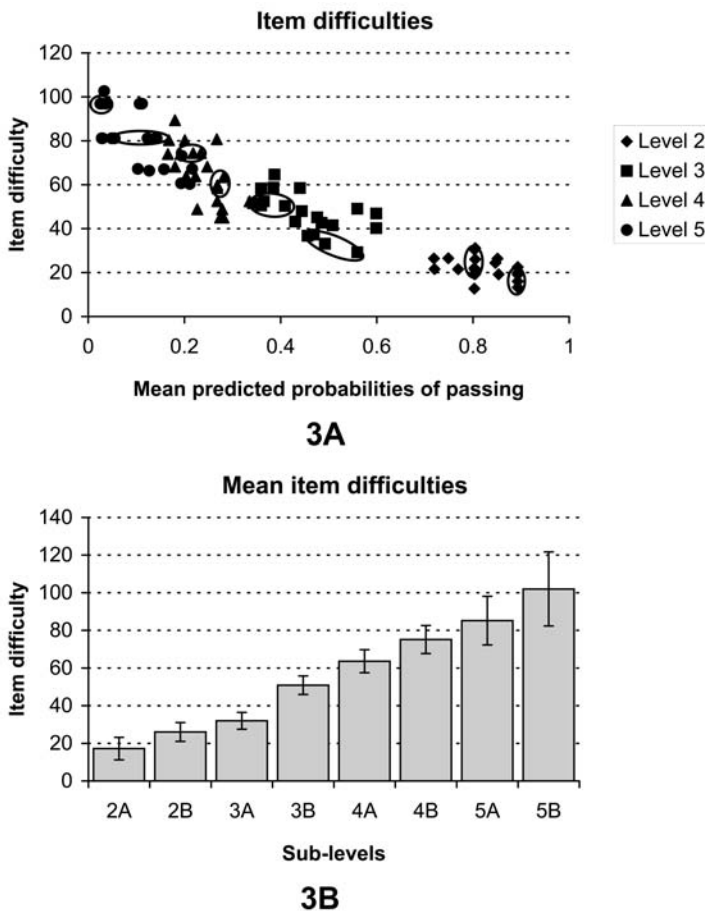


FIGURE 3 (A) Scaled item difficulties, ( $\delta_i$  from Equation 1 in Appendix 1) from the Rasch-analysis plotted on the y-axis and predicted probabilities of passing each item based on logistic regression analysis and averaged across individuals plotted in the x-axis. The items selected for further testing are circled. (3B) Means of scaled item difficulties, ( $\delta_i$  with SEM) from the selected items for further analysis. Sub-levels labelled “A” are shown to be easier than the ones labelled “B” within each level.

ording to both the Rasch item difficulties and the predicted probabilities of passing (averaged over individuals) according to the final logistic regression model.

At each level, six items fitting into two sub-levels (three in each sub-level) were selected for further testing (24 items in total for levels two through five, Figure 3A). Figure 3B shows the mean difficulty measurements obtained from the Rasch-analysis for each sub-level selected for further evaluation in Experiment 2.

We performed another series of logistic regression analyses to summarize the ability of the selected 24 items (three in each sub-level) to capture variability in success rates in the experiment 1 data. To do this we compared three models with different constraints on our item difficulties parameter values (1) no parameter constraints; 24 item specific difficulties, (2) fixing item difficulties to be equal within sub-levels (i.e.,  $4 \times 2 = 8$  parameters), (3) fixing item difficulties to be equal within levels 2 to 5 (i.e., four parameters). Each model also included a subject parameter, as in the previous models. There was no significant difference between the fit of models (1) and (2) ( $\chi^2$  diff.; 4.33,  $df = 16$ , NS) indicating that the variance in item difficulties within the sub-levels is low. There was, however, a significant difference in model fit comparing the models (2) and (3) ( $\chi^2$  diff.; 36.95,  $df = 4$ ,  $p < .001$ ). This shows that using the sub-levels explained significantly more variance than just using level.

## Discussion

In summary, the results from Experiment 1 clearly show systematic variation in item difficulties within levels. Even more importantly, the data also shows overlapping success rates between different levels, which indicates that some of the difficult items on for example level two have lower success rates compared to easy items on level three (Figures 2 and 3). This is a clear indication of the existence of other factors than number of dots contributing to level difficulty and that this effect is both large and may influence test score and interpretation of such assessments. In Experiment 1 we were also able to identify the factors contributing to these differences in difficulty and to use these findings to create a test with sub-levels. A model including the sub-levels was shown to fit the data better than one not taking sub-levels into account. At this point we have only demonstrated that the sub-levels help to distinguish children's performance using the data set on which the sub-levels were defined. To validate the progressive difficulty of our defined sub-levels we performed a second experiment. In Experiment 2, we first investigated whether the sub-levels have the same rank order of difficulty in an independent set of individuals, and, secondly, we examined the validity of our newly developed test with sub-levels, by comparing it to a commonly used VSWM test from the AWMA battery, the Odd One Out (OOO)-test. This test was chosen for its similar administration procedure (computerized and requiring pointing responses) and for its task demands including a distinct manipulation component. The test-retest reliability of the OOO is 0.81 for children aged 4.5 to 11.5 years (Alloway et al., 2006).

## EXPERIMENT 2

### Method

#### *Subjects*

The subjects were recruited through flyers and preschools in the Stockholm area and informed consent was obtained from all of the parents of participating subjects. A total of 27 children aged 4

with a mean age of 50 months, ( $SD = 2.1$  range 47–55 months, 9 girls) and 33 children aged 6–7 with a mean age of 83 months ( $SD = 3.1$ , range 77–89 months, 13 girls) participated in the study. All of the children were Caucasian and none of the children had any neurological or psychiatric diagnosis as determined with teacher reports.

### Tasks

The tasks used in Experiment 2 were the GT described in Experiment 1 and the visuospatial task OOO (AWMA, Harcourt Assessment, Packiam Alloway, 2007). The 24 items tested on the GT were the ones selected from the analysis of the results from Experiment 1 (sub-levels shown in Figure 3B). The test was terminated after six subsequent failed items. The OOO-test is computerized and consists of two concurrent tasks, one perceptual discrimination task and one visuospatial memory task. The child is presented with three squares each containing a shape. Two of the shapes are identical, and the child has to make the immediate judgment and respond to which one is the odd one out. After that, three empty squares appear in which the child has to indicate where the odd shape had previously appeared. If the child passes this level he or she will then face two subsequent sets of shapes, after which the child has to point to the screen and indicate where the odd shapes appeared in the correct order. The numbers of sets become increasingly longer with correct performance. The test is terminated after three incorrect responses out of six possible on a level (defined as the same number of sets of shapes). There are two scores obtained, one precision score (for identifying the odd one out) and a memory score (for remembering the position of the identified odd shape). Only the memory score will be used here.

### Procedure

The subjects were tested in a separate and quiet room at their preschools. All of the children performed the same items until the termination rules described earlier were applied. All of the subjects were tested at one occasion and in general there were no breaks required (total testing time ranging from approximately 10 to 25 min for the two tests). One 6-year-old failed to complete the OOO due to technical failure.

### Results

The results from the Rasch analysis conducted on the raw data from Experiment 2 supported the findings from Experiment 1 by showing sub-levels that were appropriately spread along the continuum of increasing difficulty. The sub-levels were neither overlapping within levels (consisting of the same number of dots) nor between them. In order to increase the power of the analysis, data from Experiment 1 and 2 were collapsed into yet another Rasch analysis. The item fit statistics showed one item without variability in responses and one item showing a high outfit value (a closer look indicating that this item acquired more correct answers than expected), both kept in the analysis due to the normal Z values. The person separation reliability for the Rasch analysis on the data from Experiment 1 and 2 was 0.86, the item separation reliability 0.98 and Cronbach's alpha was 0.82. Thus, the test showed adequate separation ability and high reliability.

To confirm the results of the systematic variance within and between levels in Experiment 1, the same series of logistic regression models were run on the data from Experiment 2. Similarly as

in Experiment 1, we tested models with all 24 items (1), the sub-levels (2) or the levels (3) as the parameters. As before, each model also included a subject parameter. There was no significant difference between model (1) and model (2),  $\chi^2$  diff.; 18.55,  $df = 16$ , NS, but model (2) showed a significantly better fit compared to model (3),  $\chi^2$  diff.; 52.61,  $df = 4$ ,  $p < .001$ .

To study the validity of the GT, correlations were studied between the total sum of correct scores on the GT and the OOO-test. The results showed that these two tests were highly correlated ( $r = 0.83$ ).

## Discussion

Experiment 2 confirmed the findings from Experiment 1 in that the sub-levels were ranked in the same order in both analyses. This means that our findings of intra-level difference in item difficulties are quite robust. Further supporting this are the reliability scores obtained from the Rasch-analysis showing a reliable test. None of the items showed misfit indicating that the underlying assumption of unidimensionality of the Rasch model were met by the data. There is thus no evidence suggesting that the sub-levels are tapping different aspects of WM. The model best explaining the variance in the data was the one with the sub-levels as parameters, showing similar fit as the model with the items entered as separate parameters. This indicates that the difficulties of the items within a sub-level were quite homogenous. High correlation with another test of VSWM validated that the GT actually measures the construct we aimed to measure.

*Practical testing issues.* To address the practical testing issues we decided to shorten the test in order to decrease the testing time for these young children. We selected two items per sub-level. They were chosen based on their item difficulty measurement obtained from the Rasch analysis (from the collapsed analysis of Experiment 1 and Experiment 2) and the two most uniform items per sub-level were chosen (for patterns of the items see Appendix 2 or download Eprime script at [www.klingberglab.se/external.html](http://www.klingberglab.se/external.html)).

*Scoring method.* We propose using a scoring method similar to the one suggested by Kessels (Kessels, van Zandvoort, Postma, Kappelle, de Haan, 2000) which is the product of the highest level reached and the total number of correctly repeated patterns. We propose that the levels are adjusted according to the sub-levels and suggest assigning weights to the sub-levels in which passing items within the easy sub-level on level 2 gives 2 points and which passing an item within the easy sub-level on level 2 gives 2 points and passing an item in the more difficult sub-level gives 2.5 points and so on. To validate this scoring method, this score was correlated with the person measurement obtained from the collapsed Rasch analyses from Experiment 1 and 2 ( $r = 0.86$ ). This indicates that this scoring system represents the person's ability very well. This also shows that the removal of one item per sub-level did not affect the reliability of the test.

## GENERAL DISCUSSION

The overall aim of this study was to create a more exact VSWM test that would measure WM capacity at lower spans with greater precision. The results showed large variability in success rate of different items within each level and even overlaps in difficulty between levels. The factors that were found to contribute to this within-level variability were: the distance between dots, the num-

ber of times the imaginary line crossed itself, the number of times a dot appeared on the same axis as the previous one and the number of dots appearing along a side. Based on these results, we were then able to create a new span test with sub-levels. The sub-levels were shown to explain more variance in the data than when only having levels consisting of the same number of dots, thus providing increased precision. These findings were replicated in Experiment 2 and unidimensionality and high internal reliability were demonstrated in both Experiment 1 and 2. Finally, performance on the new test with sub-levels was shown to be highly correlated with performance on another VSWM measure, the OOO-test.

Our findings that complexities of items affect difficulty are consistent with previous research (Smirni et al., 1983). Two of the factors found to contribute to difficulty within a level were the number of crossings of the imaginary line as previously reported by Orsini et al. (2001) and Busch et al. (2005) and the distance between dots as reported by Orsini et al. (2004). Kemps (2001) showed that the difficulty effect between the structured versus unstructured paths became stronger the longer the sequence was, or perhaps, the closer they came to the individual's capacity limits. The finding of path crossings was replicated by Busch and colleagues (2005) on longer sequences (7–8 blocks) only in healthy adults. This may suggest that sub-levels are perhaps most important close to each individual's capacity limit. Experiment 1 extended previous findings regarding complexity of path configurations to a child population. In addition to the factors previously identified, we could also show that the number of times a dot appeared on the same axis as the previous one and the number of dots appearing along a side affects the difficulty of a pattern. If items within a level vary in difficulty there will be unexplained variance when performing this level. Such variance should either be eliminated or systematically identified and utilized for maximum precision, as suggested here.

One concern when creating sub-levels based on the factors presented here may be the question of whether the complexity of items are testing the same underlying ability as the test of remembering different numbers of dots/objects. Kemps (2001) showed a difficulty difference when comparing structured (continuous) versus unstructured (asymmetrical) patterns in Corsi Block tapping task. This was possibly because the structured paths contain less information than the complex ones, because part of the path can be predicted from other parts, and should therefore be easier to reproduce from memory. The easier paths should thus consume less storage capacity of the VSWM system. The question of unidimensionality in this study was addressed with the Rasch analysis. The lack of misfit supports the notion that we were measuring one dimension. The results suggested that difficulty caused by increased complexity and differences in the number of dots tap the same underlying ability.

Another aspect of this question is whether complexity and information load could be differentiated based on measurements of different parts of the WM item: encoding, maintenance, and retrieval. One previous study has found that the factor "crossings of a path" limits the memory recollection already during the encoding phase (as opposed to during rehearsal) (Parmentier & Andrés, 2006). This notion is supported by other studies addressing object complexity where it was found that objects with higher information load were associated with worse performance as compared to less complex objects (Alvarez & Cavanagh, 2004; Eng, Chen, & Jiang, 2005). This was explained by the lower visual search rate recorded for the complex objects containing more visual information (Alvarez & Cavanagh, 2004). Further support of this was presented in Eng's study (2005) showing that decrease in performance for more complex objects disappeared when encoding time was increased. For our results, this would suggest that the increased complexity of a path sequence

would be perceptually more difficult to encode, thus limiting the information reaching the memory/rehearsal phase.

While some studies show that it is the number of objects (regardless of information load) stored that defines WM capacity, even when varying presentation time and interference (Luck & Vogel, 1997; Vogel, Woodman, Luck, 2001), other studies suggest that both the complexity and the number of objects determine WM capacity (Alvarez & Cavanagh, 2004). Awh, Barton, and Vogel (2007) attempted to resolve this issue by investigating simple and complex objects using a change detection task with either a high or low sample-test similarity. They replicated the findings of Alvarez and Cavanagh (2004) and found a decrease in performance as complexity increased but also found that when sample-test similarity was high, the performance decreased regardless of complexity in the objects. Awh et al. (2007) thus suggest a two-factor model of WM claiming that capacity limits are best explained by a fixed number of active slots with limited resolution. Whether these two factors should be viewed as independent or interactive is unclear. It should be noted that these studies were all conducted with a change detection paradigm, in contrast to our study. Nevertheless, this could mean that the higher complexity in our more difficult sub-levels demands higher resolution, thus limiting the number of active slots available in WM. This could be one explanation for our data although we did not find any indications of measuring two dimensions (given the low misfit to our unidimensional model). This could also imply that the underlying capacity limitation is different when examining visuospatial configurations on a free recall task as compared to a match to sample task using complexity of objects. This is however for future studies to disclose.

We found a strong correlation between the OOO and the GT. The OOO-test requires both processing and maintenance while GT, at a superficial level at least, only requires maintenance. The OOO could thus be regarded as a WM task while the GT could be regarded as more of a short-term memory (STM) task. However, recent studies suggest that in the visuospatial domain, both STM and WM can predict general intelligence equally well (Bayliss, Jarrold, Baddeley, & Gunn, 2005; Miyake, Friedman, Rettinger, Shah, Hegarty, 2001; Unsworth & Engle, 2007). This notion is especially supported in this young age group (4–6-year-olds) suggesting that STM tasks are as demanding as WM tasks (with processing) in terms of central executive involvement (Alloway et al., 2006). Possibly, the sequential presentation of cues/dots in a span task results in later cues becoming distracters for remembering the first ones and that this contributes to the high predictive power of simple span tasks (Klingberg, 2006). The high correlation with the OOO and previous findings showing a lack of difference between visuospatial WM and STM tasks, suggests that they measure the same construct: WM.

In this study we were able to use the different factors contributing to item difficulty in order to create more precise tests for VSWM. This approach might be especially useful for measuring WM capacity in younger children, where it could be used for earlier identification of children at risk for future academic difficulties or other VSWM related problems, or to identify cognitive deficits in neurodevelopmental disorders, such as ADHD. This method could also be extended to be used for adults in order to create more sensitive tests across the whole range of capacities.

## CONCLUSION

We conclude that there is a systematic variation in difficulty due to the presence of different loads on critical factors constituting the item complexity. This variability should either be removed or,

as we suggest in this study, used to provide a more precise measure of VSWM. We propose a new VSWM test consisting of two sub-levels on each level (defined as the same number of objects). This test measures one underlying dimension, has high reliability, and is highly correlated with previously published measures of WM. We suggest using the items presented in Appendix 2 and propose a scoring system shown to have high correspondence to the person ability scores obtained from the Rasch analysis. The increased predictive power of this new test remains to be demonstrated in future studies.

## REFERENCES

- Alloway, T. P. (2007) *Automated Working Memory Assessment manual*. Harcourt, Oxford.
- Alloway, T. P., Gathercole, S. E., Willis, C., & Adams, A.-M. (2004). A structural analysis of working memory and related cognitive skills in young children. *Journal of Experimental Child Psychology*, *87*, 85–106.
- 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*(6), 1698–1716.
- Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111.
- Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, *18*(7), 622–628.
- Bayliss, D. M., Jarrold, C., Baddeley, A. D., & Gunn, D. M. (2005). The relationship between short-term memory and working memory: Complex span made simple? *Memory*, *13*(3/4), 414–421.
- Berch, D. B., Krikorian, R., & Huha, E. M. (1998). The Corsi Block-Tapping Task: Methodological and theoretical considerations. *Brain and Cognition*, *38*, 317–338.
- 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*(3), 205–228.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model. Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.
- Busch, R. M., Farrell, K., Lisdahl-Medina, K., & Krikorian, R. (2005). Corsi Block-Tapping Task performance as a function of path configuration. *Journal of Clinical and Experimental Neuropsychology*, *27*, 127–134.
- Eng, H. Y., Chen, D., & Jiang, Y. (2005). Visual working memory for simple and complex visual stimuli. *Psychonomic Bulletin & Review*, *12*(6), 1127–1133.
- Gathercole, S. E., Brown, L., & Pickering, S. J. (2003). Working memory assessments at school entry as longitudinal predictors of National Curriculum attainment levels. *Educational and Child Psychology*, *20*, 109–122.
- Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004) The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*, 177–190.
- Kemps, E. (2001). Complexity effects in visuo-spatial working memory: Implications for the role of long-term memory. *Memory*, *9*(1), 13–27.
- Kessels, R. P. C., van Zandvoort, M. J. E., Postma, A., Kappelle, L. J., & de Haan, E. H. F. (2000). The Corsi Block-Tapping Task: Standardization and normative data. *Applied Neuropsychology*, *7*(4), 252–258.
- Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Training of working memory in children with ADHD. *Journal of Clinical and Experimental Neuropsychology*, *24*(6), 781–791.
- Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., et al. (2005). Computerized training of working memory in children with ADHD—A randomized, controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, *44*(2), 177–186.
- Klingberg, T. (2006) Development of a superior frontal-intraparietal network for visuo-spatial working memory. *Neuropsychologia*, *44*(11), 2171–2177.
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.
- Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *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*, *130*(4), 621–640.
- Nichelli, F., Bulgheroni, S., & Riva, D. (2001). Developmental patterns of verbal and visuospatial spans. *Neurological Science*, *22*, 377–384.
- Orsini, A., Pasquadibisceglie, M., & Picone, L. (2001). Factors which influence the difficulty of the spatial path in Corsi's block-tapping test. *Perceptual and Motor Skills*, *92*, 732–738.
- Orsini, A., Simonetta, S., & Marmorato, M. S. (2004). Corsi's block-tapping test: Some characteristics of the spatial path which influence memory. *Perceptual and Motor Skills*, *98*, 382–388.
- Parmentier, F. B. R., Elford, G., & Maybery, M. (2005). Transitional information in spatial serial memory: Path characteristics affect recall performance. *Journal of Experimental Psychology*, *31*(3), 412–427.
- Parmentier, F. B. R., & Andrés, P. (2006). The impact of path crossing on visuo-spatial serial memory: Encoding or rehearsal effect? *The Quarterly Journal of Experimental Psychology*, *56*(11), 1867–1874.
- Schneider, W., Eschmann, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh: Psychology Software Tools, Inc.
- Smirni, P., Villardita, C., & Zappalá, G. (1983) Influence of different paths on spatial memory performance in the block-tapping test. *Journal of Clinical Neuropsychology*, *5*(4), 355–359.
- SPSS for Windows, Rel. 16.0.1. 2007. Chicago: SPSS Inc.
- Thorell, L. B., Lindqvist, S., Bergman Nutley, S., Bohlin, G., & Klingberg, T. (2009). Training and transfer effects of executive functions in preschool children. *Developmental Science*, *12*(1), 106–113.
- Unsworth, N., & Engle, R. (2007) On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin*, *133*(6), 1038–1066.
- Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 92–114.
- Westerberg, H., Hirvikoski, T., Forsberg, H., & Klingberg, T. (2004). Visuo-Spatial Working Memory Span: A sensitive measure of cognitive deficits in children with ADHD. *Child Neuropsychology*, *10*(3), 155–161.
- WINSTEPS 3.63.0 Software. (1999–2006). John M. Linacre. www.winsteps.com.

## APPENDIX 1

The data was analyzed using a one parameter item response model, the Rasch model. The Rasch model assumes that the probability of a person passing a particular test item is determined by two parameters: the ability level of the person ( $\theta_n$ ) and the difficulty level of the test item ( $\delta_i$ ). If we write  $X_{ni}$  to denote the binary outcome, where 1 = “pass,” 0 = “fail” then the probability (Pr) of passing item  $i$  is written:

$$\Pr[X_{ni}=1] = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \quad (1)$$

The assumptions of the Rasch model include measuring one underlying construct and the model gives an estimate to whether the data conforms to such uni-dimensionality.

It implies equal discrimination power among all items, which fits with the assumption that the trials measure the same construct equally well along the axis of person ability. The Rasch analysis was performed in order to obtain a linear transformation of the items' difficulties as well as the subjects' abilities on the same interval scale. In this type of analysis, test items can be ranked from easiest to hardest and the subjects ordered according to their abilities of the trait that is measured. Goodness-of-fit statistics such as mean square values and Z-values can be used to evaluate the degree of fit between the actual patterns of responses and the Rasch model. Mean square values are a ratio of the observed and the predicted residual variance and have an expected value of 1, a higher value indicating that the observed scores have greater variation than predicted (more noise), and a lower value indicating that observed scores show less variance

than expected (lacking stochasticity). For this study, the infit mean square residual value between 0.6 and 1.4 with an associated Z-value of  $\leq 2.0$  was used for acceptable goodness-of-fit. Acceptable uni-dimensionality is set to require that 95% of the items fit the Rasch model (Bond & Fox, 2001; Wright & Linacre, 1994).

The person separation reliability is an estimate of how well one can differentiate persons on the measured variable. It is the fraction of observed response variance that is reproducible (by the Rasch model) and is based on the same concept as Cronbach's  $\alpha$  but for ordinal data, and available for both persons and items. The parameters were set to generate an item measurement on the logit scale where the mean item was set to a measurement of 50 and higher scores indicated higher difficulty.

## APPENDIX 2

The Final 16 Items Organized Into Easy Sublevels (Containing an A)  
and More Difficult Sublevels (Containing a B)

### Grid task items

**2A1**

	1		
	2		

**2A2**

2	1		

**2B1**

	1		
		2	

**2B2**

			1
2			

**3A1**

	2		3
	1		

**3A2**

		2	1
3			

**3B1**

			2
		1	
	3		

**3B2**

		3	
1			
			2

**4A1**

		3	
4		1	2

**4A2**

	4	1	
	3	2	

**4B1**

		1	
2			
3			
			4

**4B2**

	2		
		1	3
4			

**5A1**

	1		
		3	
2	5		4

**5A2**

		15	
2			
3			
			4

**5B1**

1		4	
			2
3	5		

**5B2**

	1		
3			5
			4
			2