ISSN-L 1618-3169 · ISSN-Print 1618-3169 · ISSN-Online 2190-5142

Experimental Psychology

www.hogrefe.com/journals/exppsy

Edited by
C. Stahl (Editor-in-Chief)
T. Beckers · A. Bröder · A. Diederich
K. Epstude · C. Frings · M. Osman · M. Perea
K. Rothermund · S. Shaki · M.C. Steffens
S. Tremblay · F. Verbruggen

HOGREFE

Experimental Psychology

Your article has appeared in a journal published by Hogrefe Publishing.
This e-offprint is provided exclusively for the personal use of the authors. It may not be posted on a personal or institutional website or to an institutional or disciplinary repository.

If you wish to post the article to your personal or institutional website or to archive it in an institutional or disciplinary repository, please use either a pre-print or a post-print of your manuscript in accordance with the publication release for your article and our "Online Rights for Journal Articles" (www.hogrefe.com/journals).

Author's personal copy (e-offprint)

Research Article

Measuring Working Memory Is All Fun and Games

A Four-Dimensional Spatial Game Predicts Cognitive Task Performance

Sharona M. Atkins,[1] Amber M. Sprenger,[1] Gregory J. H. Colflesh,[2] Timothy L. Briner,[1] Jacob B. Buchanan,[1] Sydnee E. Chavis,[1] Sy-yu Chen,[1] Gregory L. Iannuzzi,[1] Vadim Kashtelyan,[1] Eamon Dowling,[1] J. Isaiah Harbison,[2] Donald J. Bolger,[3] Michael F. Bunting,[2] and Michael R. Dougherty[1]

[1]Department of Psychology, University of Maryland, College Park, MD, USA, [2]Center for Advanced Study of Language, University of Maryland, College Park, MD, USA, [3]Department of Human Development & Quantitative Methodology, University of Maryland, College Park, MD, USA

Abstract. We developed a novel four-dimensional spatial task called Shapebuilder and used it to predict performance on a wide variety of cognitive tasks. In six experiments, we illustrate that Shapebuilder: (1) Loads on a common factor with complex working memory (WM) span tasks and that it predicts performance on quantitative reasoning tasks and Ravens Progressive Matrices (Experiment 1), (2) Correlates well with traditional complex WM span tasks (Experiment 2), predicts performance on the conditional go/no go task (Experiment 3) and N-back (Experiment 4), and showed weak or nonsignificant correlations with the Attention Networks Task (Experiment 5), and task switching (Experiment 6). Shapebuilder shows that it exhibits minimal skew and kurtosis, and shows good reliability. We argue that Shapebuilder has many advantages over existing measures of WM, including the fact that it is largely language independent, is not prone to ceiling effects, and take less than 6 min to complete on average.

Keywords: working memory, cognitive ability, N-back, go/no-go, capacity

A recent trend in cognitive science involves characterizing how performance across a wide range of behavioral tasks covaries as a function of individual differences in working memory (WM) capacity. Since Daneman and Carpenter's (1980) seminal paper (see also Case, Kurland, & Goldberg, 1982) over 30 years ago, there has been an explosion of research linking performance on a variety of cognitive tasks to individual differences in cognitive capacity (e.g., Dougherty & Hunter, 2003; Engle, 2002; Miyake et al., 2000; Sprenger & Dougherty, 2006; Unsworth & Engle, 2005) and many more studies examining the psychometric properties and factor structure of various cognitive ability measures (cf. Conway et al., 2005; Kane et al., 2004; Oberauer, 2005). A predictable trend in this area is toward studies requiring large samples and a broader range of individual differences than may be typically observed when sampling from college campuses. The standard way of collecting the data is changing, too. In Daneman and Carpenter and probably the vast majority of studies since, experimenters administered WM tasks to participants in one-on-one testing session. Data collection could be imagined to be much more efficient if the tasks could be administered independently of direct human interaction and deployed and scored on a mass scale, such as over the internet, so as to reach a broad range of participants – not just those on a college campus – with construct validity not dependent on secondary task performance.

Many existing measures of cognitive ability, specifically WM assessments, require extensive involvement of the experimenter and are thus prone to human error and variation in performance due to differential instruction. But, there are at least five more significant challenges as well. First, WM tasks are language (or writing system) specific (i.e., the stimuli are language specific and thus only validated for that language; Sanchez et al., 2010). For instance, digit span is susceptible to differences between languages in which there are more syllables in the phonological representation for number. Welsh participants show a reduced digit span relative to English that is likely not due to inherent differences in WM capacity, but rather reflect stimulus specific properties; that is, greater number of syllables in the former versus the latter language

© 2014 Hogrefe Publishing

DOI: 10.1027/1618-3169/a000262

Author's personal copy (e-offprint)

(Ellis & Hennelly, 1980). Although it is true that materials can be translated to other languages, the process can be arduous and time consuming, and translated materials must be validated all over again. While verbal WM materials can be created for any one language, the materials are not appropriate for multilingual samples (i.e., samples that do not share the same native language). Conversely, the same would not be true for a task using visuospatial materials, in which case the materials could be used with multilingual samples, provided the instructions are given in their respective native languages.

Second, many traditional dual-task WM span tasks (a.k.a. complex span tasks) require participants to perform a secondary task that requires some kind of knowledge or skills (e.g., reading and comprehending sentences or performing algebra) while holding some memoranda in memory. This can be particularly problematic as there are substantial individual differences in these secondary tasks unrelated to the theoretical construct of WM, and participants who do not score at criterion on the processing component of the task are typically excluded from the analyses (Conway et al., 2005; Turner & Engle, 1989). For example, Unsworth, Heitz, Schrock, and Engle (2005) reported that they eliminated 15% of their participants (44 of 296) from their study due to the failure to meet criterion on the math component of the automatic-operation-span task.

Third, the laboratory WM span tasks used in research often fail to meet the APA's standards for educational and psychological testing (American Psychological Association, 1999). The minimum internal consistency reliability for measures used in experimental research is $r = .75$, but this often not the case. For example, Cowan et al. (2005) used factor analysis and structural equation modeling to differentiate two kinds of WM span tasks: those measuring the scope of attention and those measuring the control of attention. Several of the WM measures suffered poor reliability. In their Experiment 1, counting span, visual arrays, and ignored speech were among the tasks with below standard Cronbach's α (.63, .66 and .70, respectively). By comparison, Cronbach's α for digit span was above standard (.88). Marginal reliability has significant impact on correlations and data analysis outcomes. The raw correlation between counting span and digit span was .32. However, when corrected for attenuation due to poor reliability, that correlation increased to .47!

Reasons four and five for why traditional laboratory WM tasks are inherently problematic are that they are prone to ceiling effects and skew. As reported by Conway et al. (2005), perfect scores are not entirely rare. In Kane et al. (2004) for virtually every complex span task, at least some participants were at or very near ceiling, as given by the percent correct. Detecting skew in published reports is not as easy, since researchers screen data for outliers that would otherwise heavily skew results. Kane et al. (2004), for example, identified 11 of 246 participants that met criteria as outliers and replaced their extreme values with a value equivalent to the mean ±3.5 *SD*. Applying corrections to make distributions more normal is by no means unusual, but it does make it difficult to look to published reports for unadjusted measures of skewness and kurtosis. This is a topic that we will return to in our own data analysis later in this report.

More ideal measures of generalized cognitive ability should be less dependent upon interactions with human investigators, less prone to individual differences in crystalized knowledge, and applicable beyond the college campus. Further, in applied contexts researchers often face severe time constraints, so the inclusion of multiple time-intensive tasks to measure WM may be practically infeasible. In these contexts, the use of a brief WM-span task may not just be preferred, but may be all that is possible. In this paper, we report a first validation of a new measure of cognitive ability that is perfectly suited to mass testing and remote data collection. The new measure, *Shapebuilder*, is instantiated within the context of a game environment that requires participants to maintain a four-dimensional representation of a set of serially presented stimuli and recall those stimuli in sequential order. Importantly, the Shapebuilder task is easily deployable over the web and is administered and scored automatically without the intervention of a human actor. Further, that task requires very little time to administer (under 6 min on average), making it suitable for experimental or applied contexts in which researchers have severe time constraints. Herein, we describe the Shapebuilder task and then present data from six studies that substantiate its internal consistency, reliability, and convergent validity with other accepted measures of WM, while also illustrating its usefulness as a measure of WM.

Shapebuilder is a visuospatial WM task[1] in which participants are asked to remember the order and spatial serial position of a series of colored shapes. Participants view a four-by-four grid of connected squares with four shapes in four colors lining each of the four sides of the grid (see Figure 1). Each stimulus is defined by four dimensions: serial position, spatial location, shape, and color. We describe the task and scoring rule in more detail in the method section.

Although Shapebuilder entails much of the same processes as complex WM span measures such as updating, memory load, and interference resolution, it differs from the established dual-task measures of WM (cf. Conway et al., 2005) in so much as it is a singular task that truly requires participants to engage in multiple demands (stimulus dimensions) simultaneously, rather than shift back and forth between two separate (and independent) tasks. It also addresses all of the aforementioned limitations of other conventional, dual-task measures of WM (e.g., operation span, reading span, letter-number sequencing). It does not require a proctor, is not language (or writing system) specific, does not require domain-specific knowledge or skills, and has no secondary task accuracy criterion.

[1] Calling Shapebuilder a WM task is a bit presumptive at this point, since we have yet to present the data behind our claim. We could simply call it a cognitive ability task or some other more general title, but the use of too many titles may be confusing.

 © 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

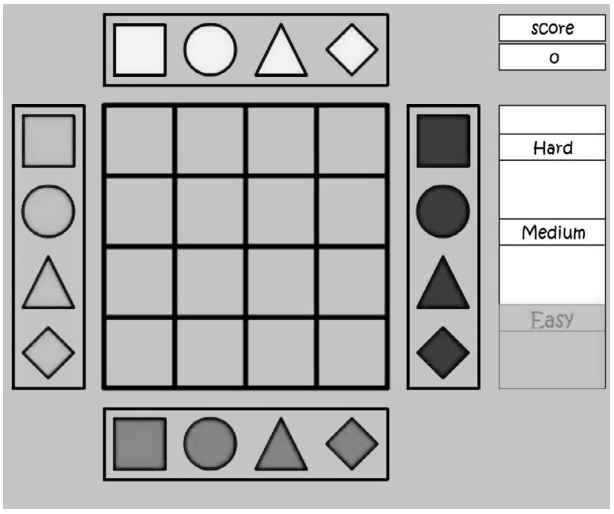

Figure 1. Screen shot of the Shapebuilder task with a gray background. Stimuli appear in the 4-by-4 grid (black). The shapes are grouped by color: yellow (top), red (bottom), blue (right), green (left). A progress bar (cyan) and score are depicted to the right of the task.

In the next section we present the results from six studies aimed at both validating Shapebuilder as a measure of cognitive ability and showing that it provides a useful tool for predicting performance on a variety of tasks that presumably engage WM and executive functioning. All of the experiments reported below were conducted in the laboratory. Although the Shapebuilder task was completed through the use of a web browser by logging into a website hosting the software, it, too, was administered in the laboratory.

Experiment 1

The goals of Experiment 1 were to establish the convergent validity and criterion validity of Shapebuilder. For convergent validity, we examined the relationship between Shapebuilder and two established measures of verbal WM, reading span and letter-number sequencing, and one visuospatial WM measure, a modified version of block span. All of these WM measures are dual-tasks with respect to the fact that they task multiple demands simultaneously. We also assess the criterion validity, by examining the relationship between Shapebuilder and other tasks for which there is typically a relationship with the WM construct to see if Shapebuilder exhibits the pattern of relationships shown with other established verbal and visuospatial WM tasks. The constructs included were: perceptual speed, abstract reasoning ability, resistance to proactive interference, inhibition, and math ability. These constructs were chosen because they have been shown to be related to WM capacity in previous research. We examined simple correlations among the tasks and also conducted an exploratory factor analysis to validate Shapebuilder as a measure of cognitive ability in relation to established WM span tasks.

Method

Participants

Participants were 117 students recruited from the University of Maryland campus and the College Park community via fliers and email announcements for participation in an experiment on WM training, which was undertaken to fulfill the requirements of an undergraduate honors program. The data reported here are for the pretest of that training study. Participants received $20 compensation for completing the pretest session. The method and results for the full training study are available online through the University of Maryland digital archives (http://drum.lib.umd.edu/handle/1903/11386). Five individuals failed to complete the pretest assessments due to time commitments. Participants ranged in age from 18 to 31 years (M = 19.53 years, SD = 2.05).

Materials and Procedure

Participants completed a battery of paper and pencil and computerized tasks during a 2.5-hr session. The tasks measured the constructs of verbal and spatial WM, perceptual speed, abstract reasoning, inhibition ability, resistance to proactive interference, and math ability, as described below. The ordering of the tasks was constant across participants as follows: letter comparison, summing to 10, canceling symbols, *g*-math, reading span, modular arithmetic, verbal learning/resistance to proactive interference, Stroop, Raven's Progressive Matrices, letter-number sequencing, number piles, block span, and Shapebuilder.

Shapebuilder

Shapebuilder is a web-administered cognitive ability task in which participants were asked to remember the order and spatial position that a series of colored shapes were presented. Participants viewed a 4×4 grid of connected squares (see Figure 1). Then, participants observed a sequence of between 2, 3, or 4 colored (red, blue, yellow, or green) shapes (circles, triangles, squares, or diamonds) that appear one at a time in one of the 16 possible grid locations. Participants were asked to remember the location of each item, the shape of each item, the color of each item, and the order that items appeared. After the final shape of a trial was presented, participants were asked to recreate the sequence by clicking on the correct colored shape and dragging it to the appropriate location. Participants completed 26 trials, of which 6 had 2 stimuli per trial, 9 had 3 stimuli per trial, and 11 had 4 stimuli per trial.

The Shapebuilder task increased in difficulty in two ways. First, trial length began at 2 and increased to 3 and then 4 items. Second, within each set of trials of a given trial length, the trials became more difficult by including

© 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

more stimuli of different colors/shapes. At the easiest level, items were all the same shape or color, but appeared in different locations. At the most difficult level, items were all different colors and shapes, and appeared in different locations. Participants received immediate feedback about the accuracy of each item; the Shapebuilder task displayed the points awarded for each item immediately after the participant released the mouse button.

The dependent variable on this task was participants' final score. Participants only received points for items that were placed in the correct location. Participants received 15 points for getting the first item of a trial correct (correct location, color, and shape) and received increasingly more points for each additional correct item: an additional 30 points for getting the second item correct after getting the first item correct, an additional 60 points for getting the third item correct after getting the first two items correct, and an additional 120 points for getting the fourth item correct after getting the first three items correct. If participants missed an item in the sequence (either entirely or partially – by forgetting one or more features), the scoring started over such that they received 15 points for the next item that was completely correct and then 30 points if the following item was correct and 60 if the following item was correct. The latter is possible only if $k = 4$. Participants received reduced points for items that were partially correct. Participants received five points for any item placed in the correct location without the correct color or shape, and 10 points for any item placed in the correct location with the correct shape, but the incorrect color. The maximum score for this task was 3,690 points.

The justification for using this specific scoring rule was twofold: (a) The score received for correctly retrieving any particular shape is monotone with realized memory load. For example, correctly retrieving the third shape in a sequence is subjectively and objectively more difficult if one has also successfully retained and retrieved the first two shapes. Because each individual shape requires memory for four dimensions (order, spatial position, color, and shape), perfect memory for the 3rd shape in a sequence requires memory for 12 total dimensions (four dimensions for the 1st shape, eight dimensions for the 2nd shape, and 12 dimensions for the 3rd shape). In contrast, when one has correctly retrieved the 3rd shape in a sequence but failed to recall the prior items, there is no way to determine what the realized memory load was, and (b) The exponential scoring rule provides motivation for participants try to remember the entire sequence of shapes, rather than focusing on one or a few shapes. Although this scoring rule may appear somewhat complicated, we provide evidence that it is both empirically justified and that it has better distributional properties than a unit scoring method, which entails assigning one point for each completely recalled shape. We present the supporting analyses on the aggregated data in the general discussion.

Established Working Memory Tasks

Reading Span (An automated version of reading span was adapted from Kane et al., 2004; Turner & Engle, 1989; Unsworth et al., 2005). Participants viewed a series of sentences and evaluated whether each sentence was grammatically correct. Participants pressed the "M" key to indicate "true" and the "Z" key to indicate "false." "T" and "F" stickers were placed on each respective key for "true" and "false." After responding to the sentence, a word was presented for 1,000 ms. If the participant failed to respond to the sentence within 2,500 ms, the program automatically advanced to presenting the word. In such cases, the response to the sentence was scored as "incorrect." Several of these sentence/word sequences were presented before participants were prompted to recall all of the words they observed in that trial in the order that the words were displayed. Responses were typed into the computer. Set sizes ranged from two to six sentence-word pairs per trial with three trials of each length, for a total of 15 trials plus 3 practice trials. Each participant saw a randomly chosen sentence paired with a randomly chosen word for each set size. The ordering of set size was constant across participants, starting with set size of 2 and then increased incrementally to a set size of 6. The dependent variable for this task was the number of correctly recalled words in the correct serial order for which the participant also correctly indicated grammaticality for the corresponding sentence in that trial.

Modified Block Span (see Atkins, 2011). In this task participants viewed a 4×4 series of squares and are asked to remember the serial order in which a sequence of yellow blocks appeared on the grid. Each block within a sequence was flashed for 1 s, one at a time, in one of the cells on the 4×4 grid. Trials were segmented into sets by the appearance of a black square mask that covered the entire grid for 1 s. After viewing a series of locations flash in a given trial, participants were asked to recall the locations that the squares were flashed in the correct order by clicking the squares in the same order that they appeared. Participants completed 16 trials of length 2–20. The task increased in difficulty by increasing the trial length, and by increasing the number of sets within each trial. For instance, for the first trial there was one set with two stimuli in the trial, and the following three trials had one set with three stimuli, then four stimuli, and then five stimuli. After this, trials were made more difficult by including increasing the number of sets, such that there were two sets of 2, 3, 4, or 5 stimuli, then three sets of 2, 3, 4, or 5 stimuli, and finally, participants viewed 4 sets of 2, 3, 4, or 5 stimuli. The dependent variable for this task was participants' score, which was computed as follows. Participants received 10 points for the first item correctly recalled, 20 for the second item in a row correctly recalled, 30 for the third item in a row correctly recalled, and so on (each additional item in a series correctly recalled given that previous items were recalled was worth 10 more points than the previous item). If an item in the series was forgotten, the scoring started over at 10 for the next item in the sequence correctly recalled. Atkins (2011) illustrated that the modified Blockspan task loaded (loading = 0.72) on a common factor with symmetry span (loading = .60), rotation span (loading = .67), and navigation span (loading = .72; see Kane et al., 2004) and reported Pearson correlations between modified block span

 © 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

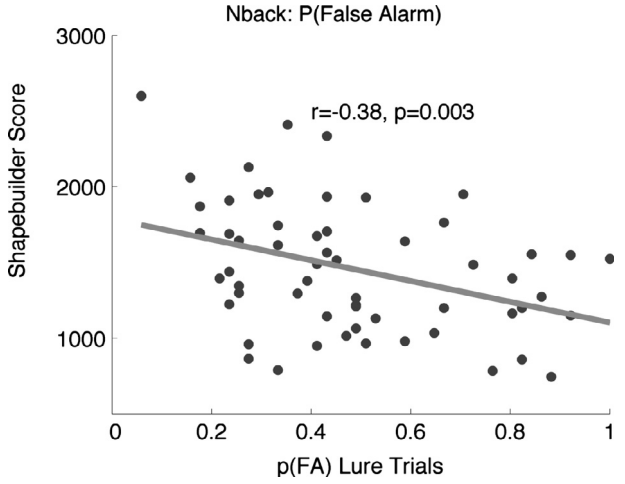

Figure 2. Scattergram showing relationship between Shapebuilder score and probability of a false alarm on lure trials for the *N*-back task in Experiment 4.

and other traditional spatial span tasks (which ranged from 0.39 for rotation span to 0.56 for navigation span) that were comparable to or higher than correlations among the traditional complex spatial span tasks (which ranged from 0.38 to 0.48). These results clearly indicated that the modified Blockspan task is a valid complex spatial WM task, as opposed to a simple spatial memory span task.

Letter-Number Sequences (LNS; Atkins 2011, Adapted from Gold, Carpenter, Randolph, Goldberg, & Weinberger, 1997; Myerson, Emery, White, & Hale, 2003). Participants viewed an alternating sequence of letters and numbers presented serially one at a time. Recall was prompted at the end of sequences. Participants were first asked to recall the numbers in ascending order and then recall the letters in forward alphabetic order (see Figure 2 below). Participants completed 14 sequences, where sequence length varied from 2 to 12 total letters and numbers. The dependent variable for this task was score. Participants received points for each correctly recalled item, but only for sequences in which all items were correctly recalled. Participants received more points for longer trials.

Perceptual Speed Tasks

The perceptual speed tasks (Canceling Symbols, Summing to 10, and Letter Comparison) were adapted from Ackerman and Cianciolo (2000). All of the tasks were administered using paper and pencil, and participants were allotted 90 s per task.

Canceling Symbols

Participants viewed a page of random letters presented without spaces. Participants were asked to search for target letters (C and D) among the distractor letters and to circle as many target letters as possible in 90 s. The dependent variable was the number of correctly circled letters.

Summing to 10

Participants viewed a page of numbers presented without spaces and were asked to circle pairs of adjacent numbers that summed to 10. Participants were asked to work as quickly as possible, as they only had 90 s for the task. The dependent variable for this task was the number of correct pairs circled in 90 s.

Letter Comparison

Participants viewed pairs of consonant-letter strings which were displayed side by side on a piece of paper. The strings were three to seven letters in length and the items of each pair were either identical or varied by one letter. Participants were asked to circle pairs that were identical. Two hundred pairs were presented, and participants circled as many identical pairs as they could in 90 s. The dependent variable from this task was the number of correctly circled pairs of items.

Inhibition Task

Stroop

Participants were presented with the names of colors in colored text (red, blue, green, or yellow). The word and font color were either congruent (i.e., the word RED presented in red font) neutral (a series of X's presented in colored font), or incongruent (i.e., the word RED presented in blue font). Participants were asked to indicate the font color as quickly as possible by pressing the key corresponding to the correct color font. The "A," "S," "K," and "L" keys on a QWERTY keyboard were labeled with stickers indicating the colors Red, Blue, Yellow, and Green respectively. The test included 144 total trials, of which 75% were congruent, 12.5% were incongruent, and 12.5% were neutral. The dependent variable for this test was the Stroop effect, defined as reaction time on correct incongruent trials minus reaction time on correct neutral trials. Thus, positive values indicate the amount of additional time to respond to incongruent trials compared to neutral trials.

Abstract Reasoning Task

Raven's Advanced Progressive Matrices (RAPM; Raven, Raven, & Court, 1998). Participants viewed eight black and white figures arranged in a 3 × 3 grid with one figure in the lower right corner missing. Participants chose the image that best completed the pattern from eight possible choices. Participants completed either the odd or even problems from Set II, which was counterbalanced across

© 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

participants. The dependent variable was the number of correctly solved problems. The maximum score for this task was 18.[2]

Proactive Interference Task

Participants were presented with a list of eight words, one at a time, with each word presented for 2,000 ms, and they were then given a mental arithmetic task in which they were shown 10 integers one at time for 1,000 ms. Participants were asked to type the sum of the digits, after which they were asked to type as many words from the list as they could remember. This basic procedure was repeated across 10 lists of words. The first three test lists consisted of words from the fruit category, the second set of three test lists consisted of words from the body part category, the third set of three lists consisted of boat-related words, and the final list consisted of words related to the category house. The second and third lists within a category comprised proactive interference (PI) trials, whereas the first list of the next block of three functioned as the release from PI list. For this task, the dependent variables were the percent correct recall for each list, the number of intrusions from prior lists, and the number of extra-list intrusions. Prior to engaging in this task participants were given a practice session that consisted of a list of weather-related words.

Math Ability Tasks

NumberPiles

Participants were asked to sum digits to a target number in a game-like environment. Participants start the game with two rows of digits inside blocks (digit-blocks) and need to click the correct number of digit-blocks to sum to the designated target number, while digit-blocks continuously fall from the top of the screen. The falling digit-blocks piled on top of the existing digit-blocks. Correctly summing the digit-blocks to the target number would cause the blocks to explode and the remaining digit-blocks to get lowered. The target was to prevent the digit-blocks from reaching the top of the screen, which would end the level. There were 10 levels of difficulty, based on speed of digit-blocks falling, number of digit-blocks to sum and whether the digit-blocks were single (1–9) or double digits (10–19). The task took 10 min. Points were awarded for every target number achieved.

G-Math

Participants observed simple arithmetic steps, displayed one at a time on the computer screen, and were asked to solve the whole arithmetic problem. Problems involved addition, subtraction, multiplication, and division operations. The answers to the problems were single digit numbers between 0 and 9. Math problems were generated randomly by the software program such that problems of different difficultly levels were equally likely. Difficulty was manipulated across trials by increasing the number of digits (between 2 and 5) for each problem (e.g., a seven operation problem involved four numbers and three arithmetic operators). Problems with 2 and 3 digits were designated as "easy" problems, whereas problems with 4 or 5 digits were labeled "hard" problems. Participants completed five practice problems and 50 test problems. The dependent variables for this task were reaction times and accuracy.

Modular arithmetic task (Beilock & Carr, 2005). The objective of modular arithmetic is to judge the truth value of problem statements (e.g., $51 \sim 19 \pmod 4$). To do this, the problem's middle number is subtracted from the first number (i.e., 51–19) and this difference is divided by the last number (i.e., 32/4). If the dividend is a whole number (as here, 8), the problem is true. If it does not equal a whole number, the problem is false. Participants completed 10 practice problems followed by two sets of 6 easy problems, two sets of six hard problems, and two sets of six medium problems, for a total of 36 total problems. For easy problems, the first two numbers were single digit numbers and the mod number was the exact difference between those numbers (e.g., $9 \sim 3 \pmod 3$). For the medium problems, the first two numbers were double digit and the mod number was a single digit number that was the exact difference between those numbers (e.g., 23–19 (mod 4)). For the hard problems, the first two numbers were double digits and the mod number was a single digit number that divided into the first difference (e.g., $28 \sim 13 \pmod 3$). There were 12 problems of each difficulty type, with half of the problems objectively true and half false. Participants responded by pressing the "M" key with their right index finger to indicate "true" and with their left index finger on the "Z" key to indicate "false." "T" and "F" stickers were placed on each respective key for "true" and "false." The dependent variables for this task were accuracy and reaction time.

Results

Of the 117 participants who initially enrolled in the study, five dropped out midway through the testing session due to time constraints, resulting in 112 participants. Additionally, data from some tasks were missing for one or more subjects due to experimenter errors or software malfunction.

Tables 1 and 2 present descriptive statistics and zero-order correlations, respectively, for the tasks in Experiment 1. Cronbach's alpha was computed based on the trial scores for the 26 individual trials of Shapebuilder and yielded a value of 0.74. This value is within the range of values often

[2] Time is the enemy of large, multicomponent studies such as this. In order to save time, we used a reduced set of 18 Raven's problems as done by Kane et al. (2004).

 © 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

Table 1. Descriptive statistics for working memory, abstract reasoning, perceptual speed, proactive interference, and math tasks

Variables	Mean	Median	*SD*	Skew	Kurtosis	*N*
1. Shapebuilder	1,581.74	1,575.00	471.56	0.16	−0.50	112
2. R-span	41.77	43.00	11.04	−0.58	0.02	115
3. Blockspan	1,467.50	1,370.00	493.52	0.33	−0.39	112
4. LNS	526.22	500.00	261.80	0.10	−0.79	111
5. RAPM	13.46	14.00	4.46	−0.31	−0.32	116
6. Stroop	165.36	141.37	227.21	1.30	11.23	115
7. PI: L1	5.68	5.75	1.03	−0.58	0.49	116
8. PI: L2	4.82	5.00	1.32	−0.21	−0.16	116
9. PI: L3	4.39	4.33	1.29	0.02	−0.42	116
10. LC	28.81	28.00	6.58	0.81	2.01	116
11. CS	41.54	41.00	12.13	2.22	12.49	117
12. ST10	22.62	23.00	5.58	−0.20	−0.15	117
13. Mod easy	1,870.72	1,739.00	626.04	1.56	4.23	111
14. Mod medium	2,904.79	2,583.00	1,441.06	2.48	9.57	111
15. Mod hard	5,229.06	4,902.18	1,964.82	1.77	5.45	111
16. GMath easy	2,011.66	1,887.78	526.24	1.38	2.08	117
17. GMath hard	5,048.89	4,798.59	1,695.84	1.68	4.29	117
18. Numberpiles	1,650.80	1,635.00	387.24	0.10	−0.39	112

Notes. R-span = reading span; LNS = Letter-Number Sequencing; RAPM = Raven's Advanced Progressive Matrices; PI:L1 = list 1 of the proactive interference task; PI:L2 = list 2 of the proactive interference task; PI:L3 = list 3 of the proactive interference task; LC = Letter Comparison task; CS = Canceling Symbols task; ST10 = Summing to 10; Mod = modular arithmetic task.

reported for traditional complex span tasks such as operation span (e.g., Conway et al., 2005).

As is evident in Table 2, Shapebuilder correlated with two previously validated measures of WM, reading span (Kane et al., 2004; Turner & Engle, 1989) and LNS (Gold et al., 1997; Myerson et al., 2003), as well as with a measure of visuospatial WM, Block span. Furthermore, Shapebuilder's pattern of correlations across the remaining tasks matched closely the pattern of the other three measures of WM. As one example, each of these tasks was positively correlated with RAPM performance, a task that measures one's ability to reason in an abstract manner; thus replicating previous research linking WM capacity and RAPM performance (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle, Tuholski, Laughlin, & Conway, 1999; Kane et al., 2004; Shelton, Elliott, Matthews, Hill, & Gouvier, 2010).

The correlations presented in Table 2 indicate that Shapebuilder is relatively well correlated with other measures of cognitive and mathematical ability. To examine this further, we performed an exploratory factor analysis using maximum likelihood estimation and a varimax rotation to determine the factor structure of various measures, and to determine the factor(s) on which Shapebuilder loads.[3] Prior to running the final factor analysis, we eliminated canceling symbols from the dataset as it created a Heywood case. Additionally, all reaction time based variables (Mod Arithmetic, GMath, and Stroop) were reverse scored by multiplying each variable by −1 so that faster responses (reflecting better performance) in the raw data were positively correlated with percent correct in the transformed data. The three different versions of the mod arithmetic were averaged to reflect a single score and the two versions of the GMath task were averaged to reflect a single score. The three verbal learning (PI) lists theoretically measure different components and were therefore entered separately into the factor analysis.

The final solution identified a 3-factor model as the best-fitting model using the proportion criterion. Moreover, this solution had a nonsignificant chi-square, $\chi^2(52) = 67.20$, and a BIC = −166.75. The BIC for the 2-factor solution (−186.59) was lower than the 3 factor solution but the corresponding chi-square statistic was significant, $\chi^2(64) = 100.66$. The chi-square difference test between the 3 and 2 factor solution revealed that the 2 factor solution provided significantly worse fit to the data, $\chi^2(12) = 33.46$, $p < .01$. The BIC for the 4-factor solution was BIC = −133.90, whereas the chi-square was nonsignificant, $\chi^2(41) = 50.37$. There was no significant difference between the 3 and 4 factor solutions according to the chi-square difference test, $\chi^2(11) = 16.83$. The final rotated factor pattern (standardized regression coefficients) for the three-factor solution and communalities are presented in Table 3. We interpret the 3-factor solution as capturing verbal memory, quantitative reasoning, and (tentatively) general WM abilities. Three tasks did not load clearly on any one factor: LNS, Stroop, and Raven's Progressive Matrices. With the exception of LNS, this is not

[3] Use of the promax rotation yielded similar factor loadings, but further revealed that the WM factor was well correlated with the quantitative reasoning factor.

© 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

Table 2. Pearson correlations for WM, abstract reasoning, inhibition, perceptual speed, and proactive interference tasks

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Shapebuilder	1																	
2. R-span	.47	1																
3. Blockspan	.62	.33	1															
4. LNS	.33	.38	.29	1														
5. RAPM	.37	.23	.29	.24	1													
6. Stroop	−.17	.02	−.02	.11	−.06	1												
7. PI: L1	.42	.40	.20	.26	.25	−.16	1											
8. PI: L2	.31	.32	.21	.06	.29	−.08	.69	1										
9. PI: L3	.29	.24	.26	.01	.28	.10	.63	.75	1									
10. LC	.06	.10	.08	.09	.09	−.07	.11	.20	.19	1								
11. CS	.15	−.08	.02	.07	−.16	−.29	.09	.01	.06	.26	1							
12. ST10	.30	.27	.37	.13	.29	−.12	.30	.32	.27	.31	.27	1						
13. Mod easy	−.43	−.35	−.32	−.32	−.36	.15	−.38	−.27	−.28	−.38	−.10	−.40	1					
14. Mod medium	−.39	−.26	−.30	−.32	−.24	.05	−.21	−.15	−.18	−.34	−.08	−.29	.82	1				
15. Mod hard	−.44	−.40	−.30	−.44	−.21	.13	−.34	−.28	−.28	−.41	−.09	−.34	.81	.79	1			
16. GMath easy	−.38	−.24	−.31	−.31	−.19	.03	−.27	−.09	−.12	−.26	−.08	−.30	.66	.60	.61	1		
17. GMath hard	−.38	−.15	−.32	−.24	−.23	.00	−.20	−.06	−.14	−.21	−.03	−.29	.58	.57	.54	.76	1	
18. Numberpiles	.48	.24	.36	.28	.30	−.12	.30	.19	.16	.27	.13	.45	−.56	−.56	−.47	−.51	−.55	1

Notes. R-span = reading span; LNS = Letter-Number Sequencing; RAPM = Raven's Advanced Progressive Matrices; PI:L1 = list 1 of the proactive interference task; PI:L2 = list 2 of the proactive interference task; PI:L3 = list 3 of the proactive interference task; LC = Letter Comparison task; CS = Canceling Symbols task; ST10 = Summing to 10; Mod = modular arithmetic task. Correlations > .19 were significant at $p < .05$.

surprising given that there are no grounds on which to suspect them to be strong indicators of the three named factors.

Within the WM factor Shapebuilder has the highest factor loading (0.85) with the other loadings being moderately lower and LNS failing to load significantly on the WM factor. Why this is the case is unclear, but one possibility is that Shapebuilder captures a more general construct than is measured by reading span and Blockspan. Partial support for the idea that Shapebuilder measures a more general construct is given by the zero-order correlations. Specifically, Shapebuilder, as opposed to the other tasks loading on the WM factor, showed the strongest correlation with the mathematical ability tasks and with Raven's Progressive Matrices, and it and reading span had approximately equivalent correlations with the verbal learning task (PI L1, PI L2, and PI L3). If Shapebuilder merely measured simple short-term memory capacity, we would not expect it to correlate so highly with measures of quantitative ability, which did not explicitly require participants to remember items for immediate recall. It should be noted that reading span and LNS loaded less strongly on the WM factor than did Shapebuilder and Blockspan. This may be reflective of the fact that Shapebuilder and Blockspan include fairly obvious spatial components, whereas LNS and reading span are more verbally oriented measures. This is evident in the zero-order correlations, which illustrate that the various measures of WM correlated with the quantitative reasoning tasks. Additionally, we performed a second factor analysis using a promax rotation, which showed that the WM factor was well correlated with the quantitative reasoning factor ($r = 0.55$) and somewhat correlated with the verbal learning factor ($r = 0.27$). The fact that the WM factor correlates so strongly with quantitative reasoning suggests that the WM factor captures something more than simple short-term memory, and replicate prior work showing a relationship between WM and quantitative abilities (Ashcraft & Kirk, 2001; Ashcraft & Krause, 2007; Beilock & DeCaro, 2007; Bull & Scerif, 2001; LeFevre, DeStefano, Coleman, & Shanahan, 2005).

Discussion

The pattern of results from Experiment 1 indicates that Shapebuilder is a valid measure of cognitive ability. Shapebuilder correlates with other measures of WM span such as reading span, LNS, and Blockspan, and it correlated with other tasks previously shown to be related to WM including math performance, proactive interference, and RAPM (Conway et al., 2002; Engle et al., 1999; Shelton et al., 2010). Tentatively, we suggest that the results of the factor analysis indicate that Shapebuilder shares construct validity with traditional measures of complex WM span.

Nominally, Shapebuilder was the single best predictor of RAPM among all of the measures included in the study. This finding, coupled with the fact that Shapebuilder, and the WM factor more broadly, was strongly correlated with mathematical ability indicate that Shapebuilder is a valid measure of WM. One limitation of Experiment 1 is that it included only one traditional measures of complex span, R-span, making it difficult to ascertain the construct validity of Shapebuilder. In the next section, we report the

 © 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

Table 3. Rotated factor pattern (standardized regression coefficients) and final communalities using Varimax rotation

	Factor			
Variables	Verbal learning	Quantitative reasoning	General working memory abilities	Communalities
Shapebuilder	.20	.23	**.85**	0.82
R-span	.24	.25	**.44**	0.32
Modified blockspan	.12	.25	**.59**	0.42
LNS	−.03	.34	.29	0.20
RAPM	.19	.18	.31	0.16
Stroop	.17	.01	.13	0.04
PI: L1	**.69**	.16	.24	0.56
PI: L2	**.94**	.13	.08	0.90
PI: L3	**.78**	.13	.10	0.64
LC	.25	**.46**	−.02	0.27
ST10	.26	**.44**	.17	0.29
Mod	.20	**.78**	.28	0.73
GMath	−.01	**.68**	.32	0.56
Numberpiles	.04	**.64**	.38	0.56

Notes. Factor loadings > .40 are in boldface. R-span = reading span; LNS = Letter-Number Sequencing; RAPM = Raven's Advanced Progressive Matrices; PI:L1 = list 1 of the proactive interference task; PI:L2 = list 2 of the proactive interference task; PI:L3 = list 3 of the proactive interference task; LC = Letter Comparison task; ST10 = Summing to 10; Mod = modular arithmetic task.

results of a second study that included a variety of complex span measures.

Experiment 2: Correlations Between Complex-Span Measures and Shapebuilder

The data presented above suggest that Shapebuilder is a valid measure of WM. Because reading span was the only measure of complex span included in that study, however, more evidence is needed to validate the convergent validity of Shapebuilder with other measures of complex WM span. To address this concern, we reanalyzed data collected as part of the first author's dissertation. In Atkins' (2011) Experiment 2, 45 participants completed a variety of cognitive ability assessments as part of an fMRI study examining the neural correlates of WM training. For our purposes, we report only a subset of the assessment measures, focusing on measures of complex span, which included Letter-Number Sequencing, Operation Span, and Symmetry Span, as well as Stroop and Raven's progressive matrices.

Method

Participants

Forty-five participants were recruited from Georgetown University and the surrounding community via the Georgetown research volunteer program and flyers placed around campus. Participants were right-handed individuals, aged 18–30 (mean age = 22.82 ± 3.81 years), native English speakers, with normal or corrected-to-normal vision, who had no personal history of neurological, neuropsychiatric, and/or psychiatric disorders and/or learning disabilities, and were not taking medication related to neuropsychiatric and/ or psychiatric disorders and or learning disabilities. Other restrictive criteria included that participants not have metal in their body, and that female participants were not pregnant, as confirmed by a pregnancy test. Participants had a mean education of 16.09 ± 1.76 years of education.

Materials and Procedure

Participants completed a battery of cognitive assessments, prior to the fmri session and were paid $20 for their time. There were two orders of task administration that were counterbalanced between participants. The administration order was such that no two WM tasks were adjacent and no two related tasks were adjacent. The assessments were all computerized, and required no interaction with the researcher beyond setting up the task. The entire cognitive battery took between 1.5 and 2 hr with breaks between assessments.

Tasks

Participants completed Shapebuilder and LNS as described in Experiment 1. The remaining tasks were as follows.

Automated Operation Span

Participants were asked to recall a series of letters. In between the presentation of the letter, they had to respond via the keyboard whether the presented solution to the math

© 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

problem is true or false. Following the keyboard response to the problem, a blank screen was presented for 500 ms, followed by a letter for 650 ms. Immediately following the letter, either another math problem appeared, or the recall cue appeared. For the recall cue, participants were presented with a letters and had to recall the letter in the serial order in which they were presented. Set sizes ranged from two to seven math problem-letter displays per trial, for a total of fifteen trials and three practice trials. Correct scores were computed by counting the total number of correctly recognized letters in the correct serial position (Unsworth et al., 2005).

Automated Symmetry Span

Participants were asked to recall the location on a 4×4 matrix of a series of red squares presented serially. In between the presentation of the red squares, they had to respond via the keyboard whether a presented image is symmetrical or not along the vertical axis. Following the keyboard response to the presented image, a blank screen was presented for 500 ms, followed by a matrix with a red square for 650 ms. Immediately following the matrix, either another image appeared, or the recall cue appeared. For the recall cue, participants were presented with a matrix and had to indicate the serial order of the location of the red block in the matrix. Set sizes ranged from two to five symmetry matrix displays per trial, for a total of twelve trials and three practice trials. Correct score was computed by counting the total number of correctly recognized arrows in the correct serial position.

Raven's Advanced Progressive Matrices

Two practice items and eighteen test items were presented to participants. Each item presented eight black and white figures arranged in a 3 by 3 grid with one figure missing. Participant chose among eight presented options the figure that best completed the pattern (Raven et al., 1998). Participants received either even or odd items at pretest and were given 18 min to complete the task.

Stroop

Participants were asked to indicate, via button press, the ink color of the series of characters presented on the screen. The series of characters was presented in Green, Blue, Red or Yellow ink, and was constructed from the words Blue, Green, Yellow, and Red for the congruent and incongruent trials, and from a series of three, four, five, or six asterisks for the baseline trials. The series of characters remained on the screen until participant response. A 750 ms fixation was presented between the character series. Participants went through a practice session of eight congruent and four baseline trials. The task consisted of 24 baseline trials, 24 incongruent trials, and 144 congruent trials. The accuracy and reaction time for the correctly identified congruent, incongruent, and baseline trials answered correctly were collected.

In addition, participants completed Mental Rotation, Posner Cueing, Mental Math, Modular Math, Grey Oral Reading Test, Word ID and Attack, Verbal Fluency and the Picture, Letter and Digit Naming, which are reported elsewhere, and are not included in this paper.

Results

As in Experiment 1, Stroop scores were computed for each participant by subtracting the RT for the Baseline trials from the RT for the incongruent trials. Scores for operation span and symmetry span were computed by summing the total number of correct items for which the corresponding math problem was also correct.

Table 4 presents the descriptive statistics and intercorrelations among the various cognitive measures. Shapebuilder scores ranged from 560 to 2,665, with a mean of 1,547.67 (SD = 495.99). While there is not enough data for factor analysis, Shapebuilder is well correlated with traditional complex span measures such as operation span (r = 0.58) and symmetry span (r = 0.48), suggesting that it measures the same construct as these two complex span measures. As well, Shapebuilder correlated with both Stroop and Ravens. Both operation span and symmetry span also correlated with Stroop. The correlation with Raven's replicates the finding from Experiment 1, whereas the fact that

Table 4. Intercorrelations among the tasks used in Experiment 2

	1	2	3	4	5	6
1. Shapebuilder	–					
2. Letter-number sequencing	.647**	–				
3. Operation span	.575**	.505**	–			
4. Symmetry span	.477**	.331*	.562**	–		
5. Ravens % correct	.369*	.366*	.107	.054	–	
6. Stroop	−.300*	−.246	−.388**	−.357*	−.049	–
Mean	1,547.67	1,090	50.59	18.38	52.35	375.11
SD	495.99	320.90	18.21	10.16	20.05	226.94
N	45	45	44	45	45	45

Note. *$p < .05$, **$p < .01$.

 © 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

Shapebuilder correlated with Stroop is consistent with prior work showing that Stroop performance is correlated with measures of complex span.

The results of Experiments 1 and 2 support the hypothesis that Shapebuilder is a valid measure of cognitive ability that shares variance with other well-established complex WM measures and measures of fluid abilities. We now turn to using Shapebuilder as an individual difference measure of WM while replicating several experimental findings within the cognitive psychology literature where WM has been shown to be a significant predictor of performance. These replications serve two purposes. First, they further validate Shapebuilder as a valid and useful measure of WM. Second, they provide an independent assessment of the replicability of the previously published effects. Our general approach was to replicate, as precisely as possible, the previously published studies that examined the relationship between complex span WM measures and other cognitive tasks, but using Shapebuilder in lieu of the previously used WM span measures.

Experiment 3: Shapebuilder and Conditional Go/No-Go

One way to show the criterion validity of Shapebuilder is to show that it relates to an outcome variable in the same manner as a known measure of WM relates to that same outcome. The Go/No-Go task is a widely used measure of motor and behavioral inhibition ability, one's ability to respond to certain cues and to refrain from responding to other cues. In this task, participants typically view a series of stimuli, such as letters, which are presented one at a time on the screen. Participants are instructed that when they see a particular cue, for instance the letter "X," they should respond (press the spacebar) as quickly as possible. For all other stimuli, participants are asked to *refrain* from acting. Thus, the task measures participants' ability to inhibit responding to distractor stimuli. The task is referenced in the developmental (Thorell, Lindqvist, Nutley, Bohlin, & Klingberg, 2009), aging (Rush, Barch, & Braver, 2006), psychopathological (Nigg, 2001), and neuroimaging (Wager et al., 2005) literatures.

Redick, Calvo, Gay, and Engle (2011) found that there was no difference between high and low WM span participants' performance on the Go/No-Go task. This result was surprising because previous research suggested that inhibition is important for WM (Hasher, Lustig, & Zacks, 2007; Hasher & Zacks, 1988; Lustig, Hasher, & May, 2001; May, Hasher, & Kane, 1999) and goal maintenance (Engle & Kane, 2004; Kane, Conway, Hambrick, & Engle, 2007), two processes that appear to be important for accurately performing the Go/No-Go task. In the Go/No-Go task, participants have to inhibit responding to distractor stimuli, and maintain the goal to respond only to relevant cues. Redick et al. (2011) argued that the result is consistent with a more recent view of WM (Unsworth & Engle, 2007a, 2007b) which argues that individual differences in WM reflect not only one's ability to actively maintain a select number of items, but also the ability to quickly retrieve information from secondary memory once activated representations have been displaced from primary memory. Redick et al. argued that the traditional Go/No-Go task does not place high demands on WM because the same cues are always linked with the same responses and because there are a minimal number of cues to maintain. Thus, the results that high and low-span participants do not differ on task performance are in line with this account of WM.

Redick et al. (2011) developed a new, conditional version of the Go/No-Go task to increase the maintenance and retrieval demands of the task and to test whether the modified task would then correlate with complex WM span measures. In this task, participants were required to respond by pressing the spacebar to the letters X and Y and to *not* respond for all other letters. Further, participants were instructed to only respond to the X or Y if the previous target was the opposite letter. For instance, if the participants saw an X as the most recent target, they were only to respond to a Y and not X's or any other letters. Indeed, in this more cognitively demanding conditional version of the Go/No-Go task, they found significant relationship between accuracy (d') and a z-composite of scores on 3 WM tasks. In Experiment 3, we examined whether Shapebuilder predicts performance on the Conditional Go/No-Go task.

Method

Participants

University of Maryland undergraduate students ($N = 60$) participated in the study for partial completion of course requirements.

Materials and Procedure

Shapebuilder

Participants completed the Shapebuilder task as described in Experiment 1.

Conditional Go/No-Go

The methods used for the Conditional Go/No-Go task were adopted from Experiment 3 in Redick et al. (2011). Participants were instructed to use the space button to make responses to the letters X and Y (Go trials), and to withhold responding to all other letters (No-Go trials). Furthermore, participants were instructed to respond only to target letters (X and Y) if the target identity switched from

© 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

Table 5. Descriptive statistics and correlations for Conditional Go/No-Go task for Experiment 2 and from Redick et al.'s (2011) Experiment 3

	Current Experiment 2, n = 60			Redick et al.'s (2011) Experiment 3, n = 171		
	M	*SD*	*r* with SB	*M*	*SD*	*r* with WM
Target	0.91	0.16	−0.13	0.95		
Distractor	0.98	0.03	0.2	0.98		
Lure	0.69	0.19	0.25	0.75		
Lag_0 lure accuracy	0.75	0.19	0.18	0.78	0.16	0.26*
Lag_{non0} lure accuracy	0.65	0.22	.28*	0.72	0.19	.41*
Target RT mean	417	55	−0.08	405	47	−0.12
Target RT ISD	104	32	−0.12	102	27	−0.34*
d' (using only lures for FAR)	2.9	1.02	0.30*	2.7	0.98	0.45*
C (bias) (using only lures for FAR)	−0.87	0.26	0.02	−0.57	0.25	−0.01
Shapebuilder	1,531.08	413.96				

Notes. RT = response time; SB = Shapebuilder score; ISD = individual standard deviations. RT analyses only conducted on correct Go trials. In Redick et al. (2011, WM was scored as a *z*-composite of Operation-span, Symmetry-span, and Running-letter span. *$p < .05$.

the previous target identity. For instance, in the sequence "X L P X Y," participants should press the spacebar for the first X in the sequence, NOT for the L (a distractor item), NOT for the P (a distractor item), NOT for the X (now a lure item), and YES for the Y (a target item). Distractor trials made up 50% of the trials, target items comprised 40% of trials, and lure items comprised 10% of trials.

Letters were presented for 300 ms followed by a blank screen for 700 ms. Participants had a total of 1,000 ms to respond to each letter. Letters were white and were presented one at a time on a black screen. Participants performed a practice block consisting of 40 stimuli (20 distractor, 16 target, and 4 lure trials). During practice participants were given visual, verbal feedback for any errors they made. Then participants completed 3 blocks of experimental trials, and each block had 200 trials (100 distractor, 80 target, and 20 lure trials). Participants completed the entire task in approximately 14 min.

Results and Discussion

The left half of Table 5 presents the results from conditional go/no go task. Participants had high levels of accuracy for correctly responding to target and distractor stimuli, but had lower accuracy for lure items, $F(2, 116) = 16.78$, $p < .05$. Mean Go stimulus RT was 417 ms, with *M* variability in response times for Go stimuli = 104 ms. D' for target and lure trials was 2.92 ($SD = 1.02$), with a mean bias of C = −0.87 ($SD = 0.25$).

Shapebuilder scores ranged from 625 to 2,020, with $M = 1{,}531.08$ and $SD = 413.96$. Participants who scored higher on Shapebuilder tended to have higher levels of accuracy on lure trials, especially when more items intervened between lures and the previous target item (lag_0 $r = 0.18$, *ns*; lag_{non0} $r = 0.28$, $p < .05$). Participants who scored higher on Shapebuilder were better at *not* responding when the task required them *not* to, especially when items were lures and when more items intervened between the lures and the most recent target. Indeed, there was a significant interaction between trial type and Shapebuilder scores, such that Shapebuilder was related more to performance on lure trials than distractor or target trials, $F(2, 116) = 4.25$, $p < .05$. Overall, Shapebuilder score was significantly related to performance on the conditional go/no-go task as measured by d' ($r = 0.30$, $p < .05$).

The right half of Table 5 presents the results from Redick et al. (2011), for comparison. The results of our Experiment 3 closely replicated the findings by Redick et al. (2011), both numerically and statistically. Similar to Redick et al.'s findings with traditional complex WM span tasks, participants who scored higher on the Shapebuilder measure were better at withholding responses to lures (false alarms) and thus had overall higher performance on the Conditional Go/No-Go task. Notably, the relationship between Shapebuilder and lure trials was only significant for lures on nonzero lag trials, as demonstrated in Figure 3. Note, too, that Redick et al. report a much stronger correlation for these trial types than they do for the lag-zero lures.

Experiment 4: Shapebuilder and the N-Back Task

Another useful task for assessing criterion validity is the N-back task. In the N-back task, participants decide whether each stimulus in a sequence matches the one that appeared N items ago. This task has been considered the "gold-standard" task in imaging studies of WM (Kane, Conway, Miura, & Colflesh, 2007; Kane & Engle, 2002).

 © 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

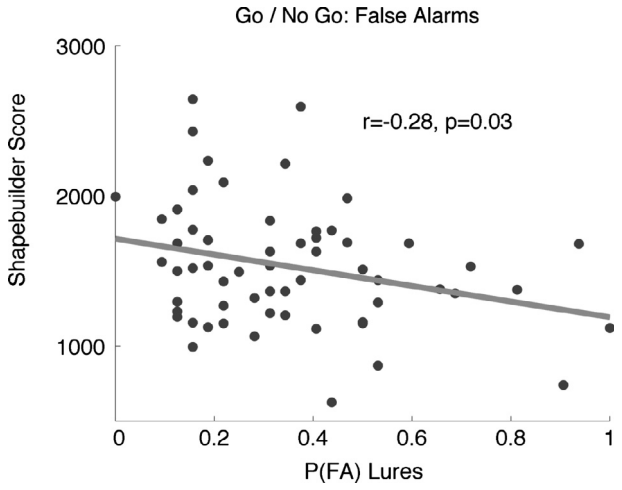

Figure 3. Scattergram showing relationship between Shapebuilder score and probability of a false alarm on lag > 0 lure trials for the conditional go/no-go task.

It consistently shows similar load effects to other WM tasks in overlapping cortical regions (dorsal-lateral prefrontal cortex and inferior parietal cortex) in PET and fMRI studies (Braver et al., 1997; Cohen et al., 1994; Petrides, Alivisatos, Meyer, & Evans, 1993; Schumacher et al., 1996; Smith, Jonides, & Koeppe, 1996) Performance on the N-back predicts individual differences in higher cognitive functions, such as fluid intelligence (Jaeggi, Buschkuehl, Perrig, & Meier, 2010; Kane, Conway, Miura, et al., 2007). Although N-back and complex WM span tasks predict fluid intelligence ability, published reports vary in the strength of the correlation between the two tasks, with some showing relatively weak correlations (Jaeggi et al., 2010; Kane, Conway, Miura, et al., 2007). Further, they account for independent variance in fluid intelligence (Kane, Conway, Miura, et al., 2007). Thus, our purpose here is primarily to assess the criterion validity of Shapebuilder for predicting N-back, though if Shapebuilder behaves like other standard complex span tasks, then we would expect to see a similarly low correlation with N-back.

Method

Participants

University of Maryland undergraduate students ($N = 58$) participated in the study for partial completion of course requirements.

Materials and Procedure

Shapebuilder

Each participant completed the Shapebuilder task as described in Experiment 1.

N-back

The stimuli for the N-back task included upper- and lower-case letters from the English alphabet. Letters were presented one at a time. Each letter was displayed for 500 ms followed by an interstimulus interval of 2,000 ms, after which the next letter was displayed. Participants were instructed to respond by pressing the "1" key on the number keypad if the current letter matched the letter presented N letters ago or "2" if the letter did NOT match the letter presented *N* letters ago. Participants were instructed to treat upper- and lower-case versions of a letter as the same letter. Participants were shown several examples for different levels of *N* before beginning the task. In the task, participants completed 50 trials at each of three levels of *N*: 2, 4, and 6. Participants first completed all 50 $N = 2$ level trials, then completed the $N = 4$ trials, and completed the $N = 6$ trials last. For the $N = 2$ trials, there were 11 targets, 17 lures, and 22 distractors. For the $N = 4$ trials, there were 11 targets, 16 lures, and 23 distractors. For the $N = 6$ trials, there were 9 targets, 17 lures, and 24 distractor items.

Results and Discussion

Table 6 presents descriptive statistics for *N*-back performance. Generally, performance on lure and distractor items remained constant across *N* level; however, performance on target items became significantly worse as *N* increased signified by a *N*-level by item type interaction for percent correct responses, $F(4, 54) = 14.87$, $p < .05$.

Shapebuilder scores ranged from 745 to 2,600, with $M = 1{,}459.57$ and $SD = 426.13$. Performance on the *N*-back task was related to Shapebuilder scores (see Table 7). D' scores were significantly correlated with Shapebuilder scores for each level of *N* as well as overall. Further, Shapebuilder significantly predicted D' performance across all levels of *N*, $F(1, 56) = 7.74$, $p < .05$, again suggesting that it is a valid measure of higher-level cognitive functioning. Looking at only false alarms, people who scored higher on Shapebuilder produced fewer false alarms ($r = -0.38$, $p < .05$), as demonstrated in Figure 2.

Table 6. Mean (*SD*) N-back Performance in Experiment 4

	Target %C	Lure %C	Distractor %C	D Prime	C Bias
$N = 2$	0.57 (0.28)	0.52 (0.29)	0.56 (0.28)	0.24 (2.01)	−0.09 (0.66)
$N = 4$	0.33 (0.21)	0.52 (0.26)	0.57 (0.29)	−0.66 (1.70)	0.35 (0.61)
$N = 6$	0.30 (0.19)	0.54 (0.25)	0.56 (0.28)	−0.82 (1.77)	0.39 (0.59)
Overall				−0.24 (1.17)	0.17 (0.36)

© 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

Table 7. Correlations between *N*-back performance and Shapebuilder in Experiment 4

N-Level	Measure	Shapebuilder
N = 2	*D′*	0.34*
	C bias	0.07
N = 4	*D′*	0.28*
	C bias	0.16
N = 6	*D′*	0.27*
	C bias	0.20
Overall	*D′*	0.32*

Note. *$p < .05$.

Experiment 5: Shapebuilder and the Attentional Networks Task

The Attentional Networks Test (ANT) was developed to quantify people's performance on three attentional components: orienting, alerting, and executive attention (Fan, McCandliss, Sommer, Raz, & Posner, 2002). The alerting component of attention helps people maintain an alert state. Orienting represents the ability to select information from sensory input. Executive control helps people resolve conflict among responses.

The ANT is a combination of the cued reaction time task (Posner, 1980) and the flanker task (Eriksen & Eriksen, 1974). The ANT requires participants to determine whether a central arrow points left or right. The arrow appears above or below a fixation point and may or may not be accompanied by flankers. The efficiency of the three attentional networks is assessed by measuring how people's reaction times are influenced by alerting cues, spatial cues, and flankers.

Because WM capacity is assumed to capture one's ability to select goal-relevant information and ignore potential distraction, researchers have hypothesized and found that differences in WM relate to performance on the Flanker task (Heitz & Engle, 2007). In fact, Redick and Engle (2006) found that high and low-span participants (using an extreme-groups design) differed in the executive score component of the ANT task. Assuming that Shapebuilder measures aspects of cognitive ability shared by complex span measures, we predicted that decrements in performance on the incongruent flanker trials in comparison to congruent flanker trials on the ANT task (captured in the executive score) would be related to participants' performance on the Shapebuilder task.

Method

Participants

University of Maryland undergraduate students ($N = 59$) participated in the study for partial completion of course requirements.

Materials and Procedure

Shapebuilder

Participants completed the Shapebuilder task, as described in Experiment 1.

ANT

Participants were presented with a series of trials in which they viewed five symbols (arrows or straight lines), and for each display they were asked to determine whether the middle arrow was pointing right or left. Participants first viewed a fixation point, indicated by a + symbol, then saw one of four possible cues, and then saw the target display (see Fan et al., 2002). The target display always appeared in one of two locations: either directly above or below the fixation point. Participants either saw no cue, a central cue (an asterisk appearing at the location of the fixation point), a double cue (an asterisk appearing at each of the two possible locations of the target), or a spatial cue (a single asterisk occurring at the same location that the target would ultimately occur. Further, there were three possible Flanker types. The central arrow either was shown with no arrows but rather straight lines on either side of it (neutral condition); with arrows pointing in the same direction as the target arrow (congruent condition); or with arrows pointing in the opposite direction as the target arrow (incongruent condition).

Participants were asked to focus their attention on the fixation point and then press the right arrow button on the keyboard if the central arrow pointed right, or to press the left arrow button on the keyboard if the central arrow pointed left. Participants were asked to respond as quickly and accurately as possible. Participants completed a practice block of 24 trials (with feedback) and three test blocks of 96 trials each with no feedback. The entire task took approximately 25 min to complete.

The alerting effect was calculated by subtracting the mean RT of the double-cue conditions from the mean RT of the no-cue conditions. When no cue is presented, attention tends to be spread across both possible cue locations. The double cue keeps attention spread in these two locations, but provides temporal information that the target will appear very soon. The orienting effect was calculated by subtracting the mean RT of the spatial cue conditions from the mean RT of the center cue. The executive control effect was calculated by subtracting the mean RT of all congruent flanking trials (across all cue types) from the mean RT of incongruent flanking trials.

Results and Discussion

We found a mean alerting effect of 43 ms ($SD = 29$), a mean orienting effect of 48 ms ($SD = 36$), and a mean executive control effect of 112 ms ($SD = 40$). Table 8 presents RTs for ANT trials separated by Cue and Flanker type. Participants responded faster and with greater accuracy for congruent and neutral trials than for incongruent trials

 © 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

Table 8. Reaction time for the ANT task as a function of Cue and Flanker type in Experiment 5

	Cue			
Flanker	Double	None	Center	Up/Down
Congruent	497 (81)	541 (74)	514 (94)	474 (83)
Incongruent	617 (96)	647 (91)	637 (113)	571 (89)
Neutral	485 (71)	538 (76)	500 (76)	461 (64)

Note. Standard deviations in parentheses.

Table 9. ANT accuracy as a function of Cue and Flanker type in Experiment 5

	Cue			
Flanker	Double	None	Center	Up/Down
Congruent	0.90 (.06)	0.90 (.07)	0.91 (.05)	0.90 (.07)
Incongruent	0.82 (.13)	0.83 (.12)	0.82 (.13)	0.85 (.13)
Neutral	0.90 (.05)	0.90 (.07)	0.90 (.06)	0.90 (.05)

Note. Standard deviations in parentheses.

(Main Effect Flanker Type: $F(2, 55) = 43.67$, $p < .05$). Further, participants responded faster when cues were present than when cues were absent, with the fastest RTs when the cue provided location information (Main effect Cue Type: $F(3, 54) = 13.18$, $p < .05$). There was no interaction between Cue and Flanker type on reaction times.

Shapebuilder scores ranged from 530 to 2,875, with $M = 1{,}507.67.57$ and $SD = 564.53$. Shapebuilder score significantly predicted reaction time, $F(1, 56) = 12.60$, $p < .05$. However, Shapebuilder score did not interact with Flanker type or Cue Type in predicting RTs.

Table 9 presents mean accuracy as a function of Cue type and Flanker type. Participants were more accurate for congruent and neutral trials than for incongruent trials (Main Effect Flanker type: $F(2, 55) = 8.44$, $p < .05$). However, there was neither a Main effect of Cue type on accuracy, nor interactions between cue type and Flanker Type. Further, Shapebuilder Score did not predict accuracy, $F(1, 56) = 2.99$, $p > .05$, and did not interact with Cue Type or Flanker Type (p's $> .05$).

Shapebuilder score did not correlate significantly with any of the three ANT components (Alerting score $r = 0.09$, $p > .05$; Orienting Score $r = -0.01$, $p > .05$; Executive Score: $r = -0.09$, $p > .05$). Overall, Shapebuilder scores are generally related to the speed at which participants respond on the ANT, but did not relate in a significant way to participants' ability to ignore irrelevant information on incongruent Flanker trials in comparison with Congruent or Neutral trials. These surprising results will be taken up in the General Discussion section.

Experiment 6: Shapebuilder and Task Switching

Discriminant validity, an important aspect of construct validity, tests whether measures that are supposed to be unrelated are, in fact, unrelated. This brings us to task switching which, though an important cognitive process, has shown little relationship to measures of complex WM. Task switching is assumed to reflect ones ability to flexibly shift attention from one activity to another without making errors or having large delays. People have higher error rates and are slower when switching from one task to another than when completing the same, repeated task (Liefooghe, Barrouillet, Vandierendonck, & Camos, 2008; Meiran, 1996; Rogers & Monsell, 1995). Some suggest that task switching is a main process of the WM system (Barrouillet, Bernardin, & Camos, 2004; Cowan, 2005). Others have shown that task-switching costs relate inversely to WM capacity (Liefooghe et al., 2008). Schneider and Logan (2005) suggested that while executive control may not be required during each trial of task-switching tasks, it may still be necessary for generally focusing attention to complete the task without frequent top-down interventions. However, Kane, Conway, Hambrick, et al. (2007) reviewed a series of studies examining the relationship between WM capacity and task-switching costs and found no relationship. Kane et al. argued that task switching may not require executive attention ability, but rather that task-switching costs result from priming effects. Furthermore, in analysis latent variable study, Miyake et al. (2000) found that task switching was not related to the latent factor underlying performance on complex span tasks, and Friedman et al. (2006) found no link between task switching and measures of general fluid intelligence.

Based on previous work examining task switching, WM, and higher-order cognition, we hypothesized that task-switching costs would not be related to performance on Shapebuilder. To test this, we conducted a study in which we measured Shapebuilder performance and had participants complete a task in which they were cued before every trial to do the "high-low" task or the "even-odd" task. Furthermore, we manipulated (within-participants) whether the same cue was used to represent the task, or whether the cue switched ("high-low" vs. "magnitude"). The latter manipulation was done so that we could estimate both the effect that switching tasks had on performance when trials switched from estimating magnitude to estimating parity (odd vs. even), as well as cue switching costs when trials maintained the same task but used different cues to indicate the task to perform ("high-low" vs. "magnitude"). Schneider and Logan (2011) argued the importance of teasing apart task-switching costs from cue-switching costs.

Method

Participants

University of Maryland undergraduate students ($N = 62$) participated in the study for partial completion of course requirements. Seven participants did not complete the Shapebuilder task and were omitted from the analyses, leaving 55 participants.

© 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

Materials and Procedure

Shapebuilder

Participants completed the Shapebuilder task, as described in Experiment 1.

Task Switching

Participants were asked to make magnitude (lower/higher than 5) and parity (odd/even) judgments of target digits (1–9, excluding 5). The words *Magnitude* and *Low-High* cued the magnitude task and the words *Parity* and *Odd-Even* cued the parity task.

Each trial in a block began with a 500 ms fixation display. A cue was then presented centrally, replacing the fixation display. After a Cue-Target Interval (0, 100, 200, 400, or 800 ms), a target was presented. The Cue and target remained visible until participants made a response, and then the screen was cleared for 500 ms. The next trial commenced immediately thereafter. The responses were made with the Z and slash keys on a QWERTY keyboard, with same-task categories assigned to different keys and category response assignments counterbalanced across participants. Reminders of the category-response assignments appeared in the bottom corners of the screen during the experiment. Participants were instructed to respond quickly and accurately. Participants completed one practice block with 62 trials and one main block with 200 trials. Cued trials were randomly selected from the full set of Cue × Target × CTI combinations. Several types of trials existed: Cue repetition trials in which the same cue occurred for the previous and current trial. Task repetition trials in which the Cue word switched from the previous trial to the current trial, but the Task remained the same (e.g., "Magnitude" for trial 1 and "High-Low" for trial 2). Task-Switching trials in which participants were asked to complete two different tasks on two successive trials (e.g., "Magnitude" for trial 1 and "Odd-Even" for trial 2).

Results and Discussion

Overall, participants made quicker responses when the Cue-Target Interval was longer, $F(4, 58) = 54.76$, $p < .05$, and when performing cue-repeat trials (vs. task or cue-switching trials), $F(2, 60) = 5.64$, $p < .05$. There was also an interaction between cue/task type and CTI for response time, $F(8, 54) = 2.79$, $p < .05$; see Figure 4) Similarly, participants had higher error rates when performing cue and task-switching trials in comparison with cue repetition trials, $F(2, 60) = 25.11$, $p < .05$. However, participants' error rates were not significantly affected by the CTI, $F(4, 58) = 2.16$, *ns*, and there was no interaction between Cue/Task type and CTI, $F(8, 54) = 0.76$, *ns;* see Figure 5.

Shapebuilder scores ranged from 740 to 2,645 ms, with $M = 1{,}433.82$ ($SD = 433.14$), and were largely unrelated to

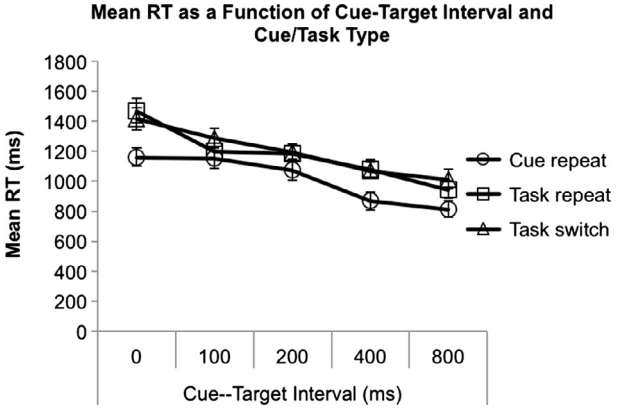

Figure 4. Mean reaction times (RT) as a function of cue-target interval and cue/task type. Error bars show standard errors.

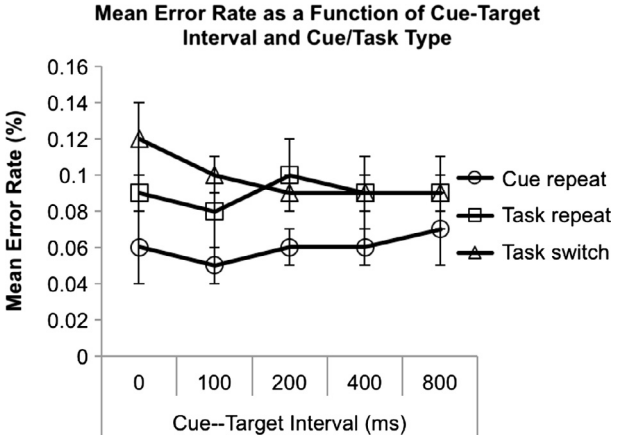

Figure 5. Mean error rates as a function of cue-target interval and cue/task type. Error bars show standard errors.

any of the reaction time variables on task-switching task (see Table 10). However, Shapebuilder was a significant predictor of error rates for cue repetition trials ($r = -0.27$, $p < .05$). The magnitude of the correlations for errors on task switches and task repetitions trials was similar in magnitude, though nonsignificant.

General Discussion

We presented a new measure of WM, Shapebuilder, that can be administered without experimenter involvement, does not require knowledge of a particular language, is easily administered and deployed over the internet, and does not lead to exclusion of any participants' data. Shapebuilder correlates (i.e., has convergent validity) with previously validated measures of verbal WM (reading span and LNS), and its pattern of correlations with other tasks matches closely that of other measures of WM. All three WM measures correlated positively with Raven's

 © 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

Table 10. Mean reaction times in ms (with standard deviation, *SD*) and percent error rates (with standard deviation, *SD*) on the Shapebuilder task and correlations between each of those scores and task-switching trials, task repetition trials, cue repetition trials, task-switching cost, and cue-switching cost

	Reaction time		Error rate	
	M (*SD*)	*r* (with Shapebuilder)	*M* (*SD*)	*r* (with Shapebuilder)
Task switches	1,198 (465)	−0.07	0.10 (.10)	−0.23
Task repetitions	1,181 (434)	−0.09	0.09 (.11)	−0.22
Cue repetitions	942 (300)	−0.05	0.06 (.09)	**−0.27**
Task-switching effect	255 (229)	−0.07	0.04 (.04)	0.01
Cue-switching effect	238 (250)	−0.11	0.03 (.06)	0.01

Note. Boldface = Significance level is $p < .05$.

Progressive Matrices performance, math performance, and recall on a task of proactive interference. An exploratory factor analysis found that Shapebuilder loaded on the same factor as reading span and modified Blockspan tasks, and this factor correlated significantly with the quantitative reasoning factor. Taken together, these data suggest that Shapebuilder measures important cognitive functions that are necessary for higher-level processing.

The hypothesis that Shapebuilder measures WM capacity is further supported by the results of Experiments 2, 3, and 4, and to a lessor extent Experiment 6. These experiments showed the convergent and criterion validity of Shapebuilder. Experiment 2 showed that Shapebuilder was well correlated with two widely used measures of complex span, operation span, and symmetry span (r's > 0.47), as well as LNS, Raven's, and Stroop. In Experiment 3, we found that Shapebuilder performance was related to performance on the Conditional Go/No-Go task, replicating findings from Redick et al. (2011). In this task, like the traditional Go/No-Go task, participants were asked to respond on some trials and not on others, based on prespecified cues (X and Y for "Go" and all other letters for "No-Go"). However, participants were asked to only respond on trials when the cue alternated from the previous "Go" cue (only respond to "X" if the previous Go-cue was "Y"). This task required participants to quickly retrieve information from secondary memory once activated representations were displaced (i.e., the previous cue identity). Furthermore, this task required participants to overcome interference when viewing lure items (i.e., when viewing an "X" but the previous Go-cue was also an "X"). Participants who scored higher on the Shapebuilder task tended to have higher levels of accuracy on lure trials, especially when more items intervened between lures and the previous target item. Redick et al. (2011) found similar results using complex WM span measures of cognitive ability.

In Experiment 4, we examined the relationship between Shapebuilder performance and performance on the *N*-back task, a task known to predict individual differences in fluid intelligence (Jaeggi et al., 2010; Kane, Conway, Miura, et al., 2007). Studies have found that *N*-back and complex-WM span tasks have only weak correlations with each other. Shapebuilder was related to D' on the N-back task overall, and individually for *N*-back at levels 2, 4, and 6, at moderate levels, again showing criterion validity.

In Experiment 6 we examined the relationship between Shapebuilder performance and task switching. Although there were no significant relationships between Shapebuilder and switch costs, we did find a significant, albeit weak, relationship for error rates on cue repetition trials. The failure to find a relationship with switch costs is not surprising, as Shapebuilder is not a dual-task and does not have a task-switching component. As it is, the link between task-switching ability and complex WM is inconsistent and it is not presently believed that task switching has a substantial relationship to WM (Kane, Conway, Hambrick, et al., 2007; Miyake et al., 2000; Oberauer, Süß, Wilhelm, & Wittman, 2003; Unsworth & Engle, 2007a).

Perhaps the most surprising result was the failure to replicate the relationship between performance on the ANT and WM capacity (Experiment 5). Previous research linked complex WM span tasks to performance on the Flanker task (Heitz & Engle, 2007), and even directly to performance on the executive score of the ANT task (Redick & Engle, 2006). As such, we hypothesized that Shapebuilder would predict performance on the executive score of the ANT task. This was clearly not the case. There are a number of possible reasons for our failure to replicate this result using Shapebuilder. First, it is entirely possible that Shapebuilder does not capture the specific abilities needed for performing the ANT, whereas traditional complex span tasks do. Another possibility is that we did not have sufficient variability (owing to the fact that all participants were college students), or that the failure to find the relationship was due to the inherent unreliability of the ANT (see Redick & Engle, 2006). While we cannot rule out any of these explanations entirely, data from a larger study conducted in our laboratory suggests that the failure may have been in part due to the homogenous nature of our sample. For example, in another study with 256 participants, Sprenger et al. (2013) found that Shapebuilder scores correlated weakly (but significantly) with ANT executive scores ($r = -0.20$, $p < .05$). In that study, executive scores also related weakly to performance on the Ravens task ($r = -0.13$, $p < .05$) but not at all with performance on the Reading span task ($r = 0.05$, *ns*). Thus, it appears that performance on the executive component of the ANT task

© 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

links with complex span tasks (such as Reading Span) sometimes, but not always – and when it does the relationship is rather weak. The main difference between the Sprenger et al. study and Experiment 5 was that the Sprenger et al. study utilized a community sample of varying ages (22 to 55 years), whereas Experiment 5 included only college students. Importantly, Redick and Engle (2006) used an extreme-groups design in which participants were sampled from both the community and college campuses. It's likely that the restricted range of variability inherent in a college sample (as used in Experiment 5) limited our ability detect the relationship between ANT and Shapebuilder. Note also that the relationship between complex span and ANT scores is quite weak across studies, suggesting that the tasks possibly measures different processes than do complex span tasks.

Detailed Analysis of Shapebuilder

Conditional Recall Probabilities

Aside from the empirical demonstrations provided in the six experiments, we can also examine more closely the properties of the Shapebuilder task. In particular, in our justification for the exponential scoring rule, we argued that memory for any particular item in a sequence is dependent on the requirement to hold other items in memory. If this is true, then we would expect that the probability of recalling any particular item should decrease as a function of memory load, as given by the list length. Further, we argued that memory for any particular item within a sequence would decrease as a function of serial position if participants are actively trying to maintain prior items in the sequence. These hypotheses can be empirically verified by examining the conditional probability of correctly recalling the *i*th item in a list given list length K.[4] The results of this analysis are plotted in Figure 6 for data aggregated across experiments.[5] As can be seen, recall probabilities decrease both as a function of list length and serial position. Moreover, the drop off in recall probabilities is rather dramatic and an increasing function of list length. For example, the decrease in recall rates between the first two serial positions is larger for list length 4 (0.24) than for list length 3 (0.18), which is greater than list length 2 (0.08). Analyzing recall rates from just the first two serial positions reveals a significant interaction between list length and serial position, $F(2, 346) = 104.36$, $p < .001$, as well as main effects of list length, $F(2, 346) = 247.81$, $p < .001$, and serial position, $F(1, 347) = 783.56$, $p < .001$. The same pattern holds in comparing recall rates for the first three serial positions for list lengths 3 and 4: there was a significant List Length × Serial Position interaction, $F(2, 346) = 19.20$, $p < .001$, and significant main effects of list length, $F(1, 246) = 177.03$, $p < .001$, and serial position, $F(2, 346) = 828.31$, $p < .001$. This pattern of recall rates supports the assumption that Shapebuilder becomes progressively more difficult both as a function of serial position and list length. Further, it provides partial justification for the use of a nonlinear scoring rule, an issue that we now revisit in more detail.

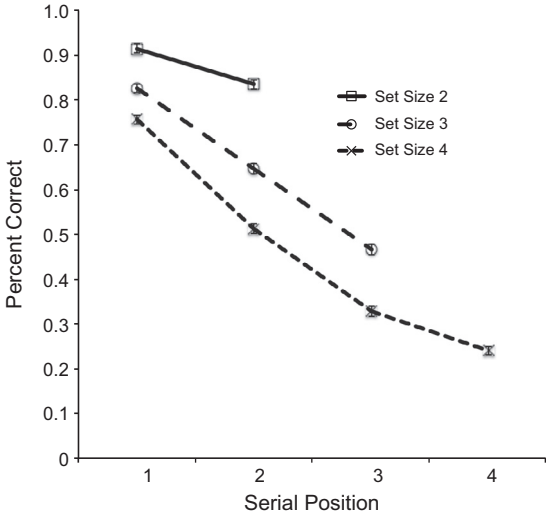

Figure 6. Results of an analysis showing the conditional probability of correctly recalling the *i*th item in a list, given list length *K*.

Analysis of the Scoring Rule

One potential criticism of our work concerns our scoring rule. Compared to most traditional complex span tasks, the scoring rule that we adopted for Shapebuilder is reasonably complicated, though we believe justified. It is important to point out that there is relatively little work investigating different scoring rules for the traditional complex span task, despite the fact that different scoring rules are used across laboratories (for a discussion of different scoring rules, see Conway et al., 2005). Thus, arguably the choice of a scoring rule is somewhat arbitrary, so long as rule yields good psychometric properties and is theoretically justified. Note that under our scoring rule, a subject receives progressively more points for correctly recalling later items within a sequence only if he or she scored perfectly on the immediately prior item or items. The decision to make the nonlinearity aspect of the scoring rule conditional on the number of previously correct items was based on the need to prevent participants from ignoring the first one or two items in a sequence, and focusing their attention on only the items that would yield the most points.

[4] To perform these analyses, the subject had to have the shape completely correct, meaning that the order, location, shape, and color all had to be correctly recalled.

[5] Note that a small number of participants ($N = 38$) participated in multiple experiments. For these participants, we only measured Shapebuilder once. Thus, the total sample size used for the analyses in this section was $N = 349$.

 © 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

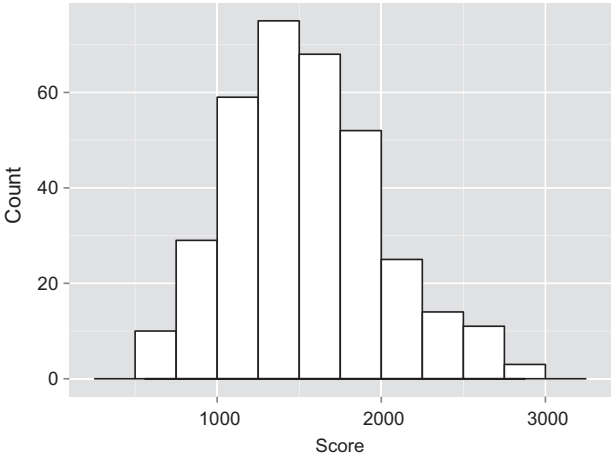

Figure 7. Distribution of Shapebuilder scores aggregated across all six studies (N = 349).

If a subject ignores the first three items in a 4-item sequence and focuses exclusively on the last item, then the effective memory load is 1. As demonstrated above, participants get progressively worse as a function of serial position and list length, suggesting that the effective memory load for each additional item in a sequence is indeed more demanding.

The use of our scoring rule is also justified based on an analysis of its distributional properties. Figure 7 plots a histogram of all of the scores across the six experiments. A few observations are worth pointing out. First, the distribution shows a small amount of positive skew (skew = 0.41, Z = 2.02, p = .04), but virtually no kurtosis (kurtosis = 2.88, Z = −0.28, p = .77). Although statistically significant, it is interesting to note that the skew is positive, as opposed to negative. As noted in the introduction, one of the limitations of existing complex span measures is that it is not unusual for participants to score at or near the maximum possible score on the task, which can lead to negative skew and ceiling effects in the measurement of WM. This implies that traditional complex span tasks lack the ability to discriminate among individuals with higher levels of ability. In contrast, no one in the experiments reported herein achieved the maximal score of 3,690 on the task (Min = 530, Max = 2,875), indicating that (a) the task is not easily amenable to ceiling effects, and (b) none of our participants could consistently and accurately maintain four shapes, which was the maximum number of shapes presented in the task. This last point is underscored by the results of a recent WM training study conducted in our laboratory, in which even participants who received extensive training (15 hr) on adaptive forms of Shapebuilder and Blockspan still could not achieve the maximum score (Sprenger et al., 2013).

Despite the fact that our scoring rule yields reasonable distributional properties, it is still useful to examine the distribution of Shapebuilder under alternative scoring metrics. Thus, we rescored the data by assigning one point for shapes that were recalled completely correctly, irrespective of whether the prior items in the sequence were correctly recalled. This is conceptually identical to the partial-credit load scoring method discussed in Conway et al. (2005). In contrast to the distribution of scores presented in Figure 7 which showed a small amount of positive skew and no kurtosis, this analysis yielded a distribution with substantial *negative* skew (skew = −0.61, z = −2.91, p < .01) and kurtosis (kurtosis = 4.08, z = 3.06, p < .01). Based on these distributional properties, we suggest that the original scoring rule is appropriate and well justified.

Administration Time, Ease of Administration, and Test-Retest Reliability

One advantage of Shapebuilder is that it takes relatively little time for participants to complete. In fact, across participants in the six studies presented here, the average completion time was 5.87 min (SD = 0.90), with a minimum of 4 min and a maximum of 10 min. In comparison, some existing complex span tasks can take up to 25 min to complete. For example, Unsworth et al. (2005) estimated completion times for the automated version of Operation span to be between 20 and 25 min.

One reason Shapebuilder takes so little time to complete is that it is a singular task that is easy to describe to participants. This contrasts with existing complex span tasks that require participants to engage in two interleaving tasks (e.g., in operation span, participants must remember a set of serially presented letters, with mental arithmetic interleaved between letters), and where participants are required to practice each task separately prior to administration. One might argue that the reliability or internal consistency of Shapebuilder would suffer as a result of its brief administration time, but this is not the case. For example, the Spearman-Brown split-half reliability was 0.756 and the Cronbach α = 0.758 across the six experiments. Further, we reanalyzed data from Sprenger et al. (2013) who used Shapebuilder as a pre/post assessment in a study of WM training. Excluding participants who trained using Shapebuilder, we calculated the test-retest correlation using the pre and post-test scores as independent administrations, which were separated by approximately 5 weeks. This correlation was r(65) = 0.82 for Shapebuilder, indicating excellent test-retest reliability. By comparison, in the same study, the test-retest reliability for an automated version of reading span and RAPM were r(71) = 0.70 and r(71) = 0.68, respectively. Note that the test-retest reliability of Shapebuilder (0.82) is on par with the test-retest reliability of the automated Operation-span (0.83), as reported by Unsworth et al. (2005), despite the fact that it requires a fraction of the time to complete.

Summary

In sum, Shapebuilder appears to be a valid measure of working memory span. As an assessment tool, it compares reasonably with traditional complex span measures, but

© 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

does not require minimal performance standards that result in data loss, which limits the generalizability to the broader population. In addition to being a valid measure, Shapebuilder also has a number of other and important qualities. For example, it takes minimal time to complete (mean time to complete = 5.87 min), is language independent, and is available via the web. We believe that this task offers a promising alternative for applied researchers who wish to collect a measure of working memory, but are otherwise limited by time or by the nature of their research designs or participant populations. For example, because our task is available via the web, it can in principle be administered to participants at home or in the workplace with minimal set up, used as a prescreening task, or even included in field studies (assuming internet capabilities). As well, because the task is language independent, it can be used in studies that cross national and language barriers. While the experiments presented herein provide promising evidence that Shapebuilder could be a useful research tool, we have only begun to explore its psychometric properties and its validity in the above-mentioned applied contexts.

Acknowledgments

This research was supported by Grant SES-0620062 awarded by the National Science Foundation to MRD and Grant N000141010605 awarded by the Office of Naval Research to MRD, DJB, JIH, and MFB. Experiment 1 was completed as part of an honors thesis at the University of Maryland submitted by TLB, JBB, SEC, SC, GLI, and VK.

References

Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied, 6*, 259–290.

American Psychological Association. (1999). *Standards for educational and psychological testing*, (2nd ed.). Washington, DC: American Educational Research Association.

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

Ashcraft, M. H., & Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review, 14*, 243–248.

Atkins, S. M. (2011). *Working memory assessment and training* (Unpublished doctoral dissertation). University of Maryland, College Park, MD (Digital Repository at University of Maryland [2012-02-17T07:03:07Z]).

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General, 133*, 83–100.

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and ''choking under pressure'' in math. *Psychological Science, 16*, 101–105.

Beilock, S. L., & DeCaro, M. S. (2007). From poor performance to success under stress: Working memory, strategy selection, and mathematical problem solving under pressure. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 33*, 983–998. doi: 10.1037/0278-7393.33.6.983

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *NeuroImage, 5*, 49–62.

Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology, 19*, 273–293.

Case, R., Kurland, M. D., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology, 33*, 386–404.

Cohen, J. D., Forman, S. D., Braver, T. S., Casey, B. J., Servan-Schreiber, D., & Noll, D. C. (1994). Activation of the prefrontal cortex in a nonspatial working memory task with functional MRI. *Human Brain Mapping, 1*, 293–304.

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D., & Minkoff, S. (2002). A latent variable analysis of working memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*, 163–183.

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

Cowan, N. (2005). *Working memory capacity*. Hove, UK: Psychology Press.

Cowan, N., Elliott, E. M., Saults, S. J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51*, 42–100.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior, 19*, 450–466.

Dougherty, M. R. P., & Hunter, J. (2003). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition, 31*, 968–982.

Ellis, N. C., & Hennelly, R. A. (1980). A bilingual word-length effect: Implications for intelligence testing and the relative ease of mental calculation in Welsh and English. *British Journal of Psychology, 71*, 43–51.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*, 19–23.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 145–199). New York, NY: Elsevier.

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.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Attention, Perception, & Psychophysics, 16*, 143–149.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience, 14*, 340–347.

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science, 17*, 172–179.

Gold, J. M., Carpenter, C., Randolph, C., Goldberg, T. E., & Weinberger, D. R. (1997). Auditory working memory and

 © 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

Wisconsin Card Sorting Test performance in Schizophrenia. *Archives of General Psychiatry, 54*, 159–165.

Hasher, L., Lustig, C., & Zacks, R. T. (2007). Inhibitory mechanisms and the control of attention. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 227–249). New York, NY: Oxford University Press.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 193–225). New York, NY: Academic Press.

Heitz, R. P., & Engle, R. W. (2007). Focusing the spotlight: Individual differences in visual attention control. *Journal of Experimental Psychology: General, 136*, 217–240.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory, 18*, 394–412.

Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 21–48). New York, NY: Oxford University Press.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the N-back task: A cautionary tale of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 615–622.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review, 9*, 637–671.

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 visuo-spatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*, 189–217.

LeFevre, J. A., DeStefano, D., Coleman, B., & Shanahan, T. (2005). Mathematical cognition and working memory. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 361–378). New York, NY: Psychology Press.

Liefooghe, B., Barrouillet, P., Vandierendonck, A., & Camos, V. (2008). Working memory costs of task switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 478–494.

Lustig, C., Hasher, L., & May, C. P. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General, 130*, 199–207.

May, C. P., Hasher, L., & Kane, M. J. (1999). The role of interference in memory span. *Memory & Cognition, 27*, 759–767.

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1423–1442.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wagner, T. (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.

Myerson, J., Emery, L., White, D. A., & Hale, S. (2003). Effects of age, domain, and processing demands on memory span: Evidence for a differential decline. *Aging, Neuropsychology, and Cognition, 10*, 20–27.

Nigg, J. T. (2001). Is ADHD a disinhibitory disorder? *Psychological Bulletin, 127*, 571–598.

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General, 134*, 368–387.

Oberauer, K., Süß, H. M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence, 31*, 167–193.

Petrides, M., Alivisatos, B., Meyer, E., & Evans, A. C. (1993). Functional activation of the human frontal cortex during the performance of verbal working memory tasks. *Proceedings of the National Academy of Sciences, 90*, 878–882.

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology, 32*, 3–25.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's progressive matrices and vocabulary scales: Section 4. The advanced progressive matrices*. San Antonio, TX: Harcourt Assessment.

Redick, T. S., Calvo, A., Gay, C. E., & Engle, R. W. (2011). Working memory capacity and Go/No-Go Task Performance: Selective effects of updating, maintenance, and inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 308–324.

Redick, T. S., & Engle, R. W. (2006). Working memory capacity and the Attention Network Test performance. *Applied Cognitive Psychology, 20*, 713–721.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*, 207–231.

Rush, B. K., Barch, D. M., & Braver, T. S. (2006). Accounting for cognitive aging: Context processing, inhibition or processing speed? *Aging, Neuropsychology, and Cognition, 13*, 588–610.

Sanchez, C. A., Wiley, J., Miura, T. K., Colflesh, G. J. H., Jensen, M. S., Ricks, T. R., & Conway, A. R. A. (2010). Assessing working memory capacity in a non-native language. *Learning and Individual Differences, 20*, 488–493.

Schneider, D. W., & Logan, G. D. (2005). Modeling task switching without switching tasks: A short-term priming account of explicitly cued performance. *Journal of Experimental Psychology: General, 134*, 343–367.

Schneider, D. W., & Logan, G. D. (2011). Task-switching performance with 1:1 and 2:1 cue-task mappings: Not so different after all. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 405–414.

Schumacher, E. H., Lauber, E., Awh, E., Jonides, J., Smith, E. E., & Koeppe, R. A. (1996). PET evidence for an amodal verbal working memory system. *NeuroImage, 3*, 79–88.

Shelton, J. T., Elliott, E. M., Matthews, R. A., Hill, B. D., & Gouvier, W. D. (2010). The relationships of working memory, secondary memory, and general fluid intelligence: Working memory is special. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 813–820.

Smith, E. E., Jonides, J., & Koeppe, R. A. (1996). Dissociating verbal and spatial working memory using PET. *Cerebral Cortex, 6*, 11–20.

Sprenger, A. M., Atkins, S. M., Bolger, D. J., Harbison, J., Novick, J. M., Chrabaszcz, J. S., ... Dougherty, M. R. (2013). Training working memory: Limits of transfer. *Intelligence, 41*, 638–663.

Sprenger, A. M., & Dougherty, M. R. (2006). Differences between probability and frequency judgments: The role of individual differences in working memory capacity. *Organizational Behavior and Human Decision Processes, 99*, 202–211.

Thorell, L. B., Lindqvist, S., Nutley, S. B., Bohlin, G., & Klingberg, T. (2009). Training and transfer effects of executive functions in preschool children. *Developmental Science, 12*, 106–113.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*, 127–154.

© 2014 Hogrefe Publishing

Author's personal copy (e-offprint)

Unsworth, N., & Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlation between operation span and Raven. *Intelligence, 33*, 67–81.

Unsworth, N., & Engle, R. W. (2007a). On the division of short-term and working memory: An examination of simple and complex spans and their relation to higher order abilities. *Psychological Bulletin, 133*, 1038–1066.

Unsworth, N., & Engle, R. W. (2007b). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*, 104–132.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*, 498–505.

Wager, T. D., Sylvester, C. Y., Lacey, S. C., Nee, D. E., Franklin, M., & Jonides, J. (2005). Common and unique components of response inhibition revealed by fMRI. *NeuroImage, 27*, 323–340.

Received July 31, 2013
Revision received January 29, 2014
Accepted February 5, 2014
Published online June 23, 2014

Michael Dougherty

Department of Psychology
University of Maryland
College Park
MD 20742
USA
E-mail mdougher@umd.edu

 © 2014 Hogrefe Publishing