

The development of a short domain-general measure of working memory capacity

Frederick L. Oswald · Samuel T. McAbee ·
Thomas S. Redick · David Z. Hambrick

Published online: 6 December 2014
© Psychonomic Society, Inc. 2014

Abstract Working memory capacity is one of the most frequently measured individual difference constructs in cognitive psychology and related fields. However, implementation of complex span and other working memory measures is generally time-consuming for administrators and examinees alike. Because researchers often must manage the tension between limited testing time and measuring numerous constructs reliably, a short and effective measure of working memory capacity would often be a major practical benefit in future research efforts. The current study developed a shortened computerized domain-general measure of working memory capacity by representatively sampling items from three existing complex working memory span tasks: operation span, reading span, and symmetry span. Using a large archival data set (Study 1, $N = 4,845$), we developed and applied a principled strategy for developing the reduced measure, based on testing a series of confirmatory factor analysis models. Adequate fit indices from these models lent support to this strategy. The resulting shortened measure was then administered to a second independent sample (Study 2, $N = 172$), demonstrating that the new measure saves roughly 15 min (30 %) of testing time on average, and even up to 25 min depending on the test-taker. On the basis of these initial promising findings, several directions for future research are discussed.

Keywords Working memory · Measurement · Complex span tasks

Working memory capacity (WMC) is among the most frequently assessed constructs in cognitive psychology and related fields (e.g., Daneman & Carpenter, 1980; Engle, 2002; Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000) and has been defined as the ability to store, recall, manage, and manipulate information in a highly active state (Engle, 2002). Under this definition, it is not surprising that WMC is related to, and highly correlated with, general factors of intelligence, in particular fluid intelligence (*Gf*), which involves storing and transforming information to solve novel and abstract problems (Ackerman, Beier, & Boyle, 2005; Beier & Ackerman, 2005; Kane, Hambrick, & Conway, 2005; Oberauer, Schulze, Wilhelm, & Süß, 2005). Measures of WMC are also correlated highly with other measures of complex cognition, such as reading comprehension and problem solving (Conway et al., 2005; Daneman & Merikle, 1996; Engle, 2002; Friedman & Miyake, 2004), as well as with measures hypothesized to reflect elementary cognitive processes, such as executive attention and executive control (Engle & Kane, 2004; Kane, Conway, Hambrick, & Engle, 2007). In terms of predicting real-world outcomes, low levels of WMC have been connected with a variety of cognitive deficits and clinical diagnoses (e.g., Attention Deficit Disorder, Alzheimer's disease, schizophrenia; Redick et al., 2012), and, more generally, WMC and other cognitive ability measures positively correlate with performance outcomes in educational settings (e.g., GPA: Colom, Escorial, Shih, & Privado, 2007), occupational settings (e.g., job performance: Schmidt & Hunter, 2004; career success: Judge, Higgins, Thoresen, & Barrick, 1999), and everyday life (e.g., health outcomes: Deary, Weiss, & Batty, 2010).

Although WMC measures are grounded in theories of cognition and are useful for predicting important personal

F. L. Oswald (✉) · S. T. McAbee
Department of Psychology, Rice University, 6100 Main Street,
MS - 25, Houston, TX 77005, USA
e-mail: foswald@rice.edu

T. S. Redick
Department of Psychological Sciences, Purdue University, Lafayette,
IN, USA

D. Z. Hambrick
Department of Psychology, Michigan State University, East Lansing,
MI, USA

and societal outcomes, from a practical standpoint, WMC measures tend to require a great deal of time to administer. Even batteries of working memory span tasks that are computerized, automated, and administered in groups can easily take 45 min for groups to complete (Unsworth, Heitz, Schrock, & Engle, 2005). Therefore, given that researchers often want to measure a wide array of psychological constructs yet are faced with a limited amount of available testing time (Stanton, Sinar, Balzer, & Smith, 2002), a shorter and psychometrically sound measure of WMC would be an asset for future research. To that end, the current study samples items across three different types of complex working memory span tasks; then by developing and applying a principled conceptual and psychometric approach to shortening these tasks and modeling the data, we create a shortened measure of WMC.

Working memory span tasks

Out of the various measures of working memory available, complex memory span tasks have received considerable research attention both within and outside the cognitive psychology literature¹ (Conway et al., 2005). Daneman and Carpenter (1980) are widely cited as having developed the first complex working memory span task, the *reading span task*, argued to measure working memory because of its presentation of both a processing component (i.e., the sentence to be read) and a storage component (i.e., the to-be-remembered word), which they posited as critical to assessing individual differences in WMC (see also Engle, Tuholski, Laughlin, & Conway, 1999). Daneman and Carpenter's version of the reading span task requires examinees to read a series of 2–6 sentences and for each sentence to make a true/false judgment about its veracity. After all sentences are judged, examinees are then asked to recall the last word of each sentence in the series. The examinee's reading span score is the highest set size (number of sentence-word pairs) that was recalled two out of three trials perfectly.

Following Daneman and Carpenter's (1980) creation of the reading span task, a series of alternative measures of complex span tasks have been developed to assess working memory capacity (for reviews of these tasks, see Conway et al., 2005; Unsworth, Redick, Heitz, Broadway, & Engle, 2009). The common requirement for each of these tasks is the pairing of a task followed by a to-be-remembered element (e.g., a letter, word, or object), such that subsequent tasks interfere with the previous elements presented (Unsworth et al., 2005). Computerized versions of several classic complex working

memory span tasks have been developed (e.g., Redick et al., 2012; Unsworth et al., 2005, 2009a, b), allowing for quicker administration times, examinee-driven administration and time limits, and automatic scoring (Redick et al., 2012). For example, in a computerized *operation span task* (Conway et al., 2005; Unsworth et al., 2005), examinees are given a series of very simple arithmetic operations. For each operation, they indicate whether it is true or false, and then they are provided with a letter to be recalled later (e.g., $2 + 7 = ?$, 9, Q). After the series is completed, examinees are then prompted with a 4×3 matrix of letters and asked to click on the letters to be remembered in the order in which they were presented. The processing (arithmetic operation), decision (true or false), storage (letter), and recall (letter matrix) phases of the automated operation span task are each presented on separate computer screens to minimize rehearsal of the to-be-remembered elements (see Redick et al., 2012).

In a subsequent study, Unsworth et al. (2009a, b) created computerized versions of the reading span and symmetry span tasks. These tasks follow the same general procedure as that presented for the automated operation span task. For the computerized *reading span* task, examinees are asked whether a presented sentence makes sense (e.g., "The prosecutor's *dish* was lost because it was not based on fact.") followed by letters to be recalled, just as in the operation span task (Unsworth et al. 2009a, b). For the computerized *symmetry span* task, examinees view an 8×8 matrix of white and black squares and determine whether the pattern is symmetrical along its vertical axis. After this judgment, examinees are presented with a 4×4 matrix of squares in which one cell is highlighted in red. After the series of matrix presentations, examinees must then recall the serial order of the positions of the red cells.

Domain-general versus domain-specific perspectives of working memory capacity

There are two major theoretical perspectives on what working memory tasks measure (Hambrick, Engle, & Kane, 2004). The *domain-specific perspective* focuses on the overlap between the specific content of the working memory task and the specific outcomes of interest. For example, Daneman and Carpenter (1980) posited that the reading span task predicted performance on a reading comprehension test, because both shared reading ability as an underlying core process. But more broadly, the domain-specific perspective is supported when convergent and discriminant validity patterns are observed, such that specific working memory tasks correlate more strongly with outcomes that require that specific ability versus those outcomes requiring different specific abilities (e.g., Daneman & Carpenter, 1980; Oberauer et al., 2000; Shah & Miyake, 1996).

¹ Although we focus on complex span tasks as measures of working memory, a number of alternative measures exist (see e.g., Cowan et al., 2005; Oberauer et al., 2000; Was, Rawson, Bailey, & Dunlosky, 2011).

By contrast, the *domain-general perspective* assumes that the processing, storage, and recall requirements are common across specific working memory tasks (not unique to each task), and it is this commonality that is primarily responsible for the correlations between working memory and relevant outcomes. Latent variable modeling operationalizes this perspective, such that a general working memory factor is supported by virtue of the reliable and relatively high correlations between various verbal and spatial working memory tasks (e.g., Engle et al., 1999; Kane et al., 2004; see also Engle, Cantor, & Carullo, 1992; Hull, Martin, Beier, Lane, & Hamilton, 2008). There is now a great deal of support for the domain-general perspective. For instance, Kane et al. (2004) reported that verbal and spatial working memory factors were highly related, sharing between 70 % and 85 % of their variance, leaving little room for reliable variance unique to each factor.

We take a balanced perspective and procedure in building a short domain-general measure of WMC: even with empirical support in the literature for the reliability and validity of the domain-general perspective, it is still important to be sensitive to the domain-specific perspective and representatively sample from tasks across numerical, verbal, and spatial modalities (i.e., operation, reading, and symmetry span tasks, respectively). This balance between empirical and substantive priorities in measurement is known as *controlled heterogeneity* in the cognitive ability domain (Humphreys, 1962), and it is a fundamental concept in psychological measurement (Little, Lindenberger, & Nesselrode, 1999).

A short domain-general measure of working memory capacity

Diverse areas of psychology are developing short psychological scales because of their many potential practical advantages. On the test-taker side, short measures may increase examinee engagement and conversely may reduce fatigue, careless responding, and attrition (Ackerman & Kanfer, 2009). On the administrator side, short measures may conserve time or free up available testing time to measure other constructs relevant to research and practice. For the cognitive abilities domain in particular, short counterparts have been derived for traditional measures such as the Raven's Advanced Progressive Matrices (e.g., Arthur & Day, 1994; Bors & Stokes, 1998) and the Weschler Adult Intelligence Scales (e.g., Jeyakumar, Warriner, Raval, & Ahmad, 2004; Miller, Streiner, & Goldberg, 1996).

We present two studies that reflect a conceptual and psychometric process for developing a viable short computerized measure of domain-general working memory.

Study 1

Data had been obtained in prior research by the third author from a large archival sample of participants who completed the computerized complex span tasks (operation, reading, and symmetry span; see Redick et al., 2012). The goal of Study 1 was to use these existing data and apply a principled psychometric measure-shortening procedure (to be described) to develop a short measure of WMC, in hopes that any tradeoffs in psychometric characteristics that come with measure-shortening would be minimal and thus recommend the use of the shortened measure in many research applications.

Method

Sample

The sample included undergraduate students from three colleges in the southeastern United States (University of North Carolina Greensboro [UNCG], University of Georgia [UGA], and Georgia Institute of Technology [GT]), as well as community-recruited adults around one of these schools (nonGT). The total dataset included 6,611 participants who completed at least one of the automated complex span tasks between 2004 and 2009. The current study used the data from 4,885 of these participants who completed all three automated complex span tasks ($n_{UNCG} = 1,258$ [25.8 %]; $n_{UGA} = 1,598$ [32.7 %]; $n_{GT} = 1,035$ [21.2 %]; $n_{nonGT} = 994$ [20.3 %]). Of these, demographic data were available for 4,445 participants who were 17–35 years of age ($M = 20.4$, $SD = 3.5$), and 39.5 % ($N = 1,757$) were male. Based on a stratified random sample across these data collection sites, the sample was split in half. We present the bulk of our results from 50 % of the sample considered the *development sample*, which is a large enough sample to consider results to be stable ($N = 2,442$). However, in the interest of replicability and to ensure that our results do not capitalize on chance, we also briefly summarize results based on the *cross-validation sample* ($N = 2,443$), where the parameter estimates established on the short WMC measure for the model in the development sample are applied to the same model in this independent cross-validation sample.

Measures

For each working memory measure—operation span, reading span, and symmetry span—participants completed a series of practice trials consisting of (1) the storage component of the task alone, (2) the processing component of the task alone, and (3) the process component followed by the storage component (set-size 2). In the actual test trials that followed, set-size orders were randomized within participants, and the time limit for each processing-storage trial was set to be equal to 2.5

standard deviations above the mean time for the processing-only responses in the participant's practice trials (see Redick et al., 2012).

Operation span Participants were presented with a set of arithmetic operations and asked to judge whether each equation was true or false (approximately half were true). After each arithmetic operation, participants were presented with an element (a letter) for recall at the end of the set. Set sizes ranged from 3–7, with three administrations for each set size (i.e., 75 total operation-storage pairs).

Reading span Participants were presented with a set of sentences of approximately 10–15 words in length and were asked to judge whether or not the sentence was sensible (approximately half were sensible). After each sentence, participants were presented with an element (a letter) for recall at the end of the set. Set sizes ranged from 3–7, with three administrations for each set size (i.e., 75 total sentence-storage pairs).

Symmetry span Participants were presented with a set of 8×8 matrices of black and white squares and asked to make a judgment as to whether the matrices were symmetrical down the vertical axis (approximately half of the matrices were symmetrical). After each matrix, participants were presented with a red square positioned in a 4×4 matrix for recall at the end of the set. Set sizes ranged from 2–5, with three administrations for each set size (i.e., 42 total symmetry-storage pairs).

Scoring For each combination of span task and set size, participants received two overall scores (see Redick et al., 2012). Participants' *absolute scores* are the number of trials in which the participant recalled all elements in the correct order without error, and participants' *partial-credit scores* incorporate error by adding up the proportions of correctly recalled elements in each trial (see Conway et al., 2005, for other alternatives). As might be expected, partial-credit scores correlated very highly ($r \geq .91$) with absolute scores, and, thus, we used participants' item-level partial-credit scores, which are generated by the program administering the measure. Table 1 presents the descriptive statistics and correlations for the partial-credit scores across the span task scores, where scores are averaged across all set sizes.

Analysis

Because set sizes for each of these complex working memory span tasks are presented to each participant in a randomized order (Redick et al., 2012), we matched participants' first, second, and third administrations of each particular combination of span and set size (hereafter referred to as an *item*),

Table 1 Means, standard deviations, and correlations for partial-credit task scores

	WM Task	Mean	SD	1.	2.	3.
1.	Operation span	57.30	13.59	.86		
2.	Reading span	53.74	15.00	.68	.89	
3.	Symmetry span	28.06	7.97	.52	.51	.80

Note. $N = 2,442$. Span scores are summed across set sizes. Alpha coefficients are in italics on the main diagonal. All correlations are statistically significant, $p < .001$

regardless of specifically where in the series the participant actually completed the trial.

Measure-shortening procedures Given this large development sample ($N = 2,442$), we were able to implement a rational procedure that shortens the working memory measure across span tasks while attempting to preserve both its substantive homogeneity and psychometric integrity. We viewed representative sampling of items across the operation, reading, and symmetry span tasks types as critical to the development of a domain-general measure of WMC (e.g., Hambrick et al., 2004; Kane et al., 2004).

The working memory and psychometrics literature informed our overall measure-reduction strategy in three ways. First, Conway et al. (2005) found that, all other things being equal in their data, working memory items with larger set sizes tended to have greater psychometric reliability. However, as a more general psychometric phenomenon, very easy items (small set sizes) can create a distribution of scores with ceiling effects, and very hard items (large set sizes) can create a distribution of scores with floor effects; therefore, the goal is to find set sizes that are in between these extreme endpoints to maximize reliable variance (i.e., accurate discrimination between people's measured levels of working memory). Keeping this in mind, we first removed the three administrations of the shortest set size items in each task (e.g., set size 2 in symmetry span, set size 3 in operation and reading span); this short measure is associated with Model 1 that we will test.

Second, even though three administrations of each set size are typical for complex span measures of WMC (likely a tradition of the procedure established by Daneman & Carpenter, 1980), fewer administrations per set size would obviously reduce the time to complete a measure (see Foster et al., 2014). We therefore removed the third administration of all set sizes to determine whether the measure shortened in this manner would retain appropriate psychometric properties. This, along with the previous revision to the measure in Model 1 (i.e., removing the shortest set sizes), results in a second short measure that is associated with Model 2 that we will test.

Finally, the largest set sizes might lead to floor effects as mentioned, and in addition, large set sizes will typically require the most time to complete. Therefore, we removed all items with the largest set size from the operation span and reading span tasks (i.e., set size 7). Combined with the two previous measure-shortening strategies (removing the smallest set sizes and removing the third administration) this results in a third shortened measure associated with Model 3 that we will test. Table 2 summarizes these three consecutive measure-reduction steps and their associated models.

Factor analyses Paralleling the measure-shortening strategy above, we tested and compared a series of confirmatory factor analysis (CFA) models for each of the progressively reduced measures and models: Models 1, 2, and 3. Models were hierarchical, with WMC being the general (second-order) factor; operation span, reading span, and symmetry span task scores being specific (first-order) factors; and composite scores (parcels) of tasks at each set-size serving as indicators of the specific factors (see Little, Cunningham, Shahar, & Widaman, 2002). There are 14 parcels: five for operation span, five for reading span, and four for symmetry span. The left panel of Fig. 1 depicts the structure of these models.

Note that some parcels were removed in Models 1 and 3 in that small and large set sizes, respectively, were eliminated; also, parcels were recomputed for Models 2 and 3 because the third administration of a given set size was eliminated. After running these CFAs (using Mplus version 6.11; Muthén & Muthén, 2010), we conducted a CFA of Model 3 that applies factor loadings derived from the development sample to the hold-out cross-validation sample of data ($N = 2,443$; see MacCallum, Roznowski, Mar, & Reith, 1994).

Results and discussion

Descriptive statistics and internal-consistency reliabilities

Table 3 presents the item-level descriptive statistics across all items and set sizes for the computerized operation, reading,

and symmetry span tasks. We also include item-remainder correlations (IRCs), where a positive correlation means that a higher score on a given item predicts a higher average score on the other two items that have the same set size (i.e., the item is correlated with a “remainder” total score that is not influenced by the item itself). A low IRC (say .10 or less) would mean that the item does not “belong” to the rest of the items that reflect the same set size and span task.

Mean scores are the proportion of people who answered correctly, and, as expected, items from shorter set sizes were easier to answer and had the highest mean scores across all tasks. The higher mean scores from these small set sizes led to a ceiling effect (range-restriction effect), creating smaller standard deviations and smaller IRCs across span tasks; this ceiling effect, in turn, is related to high values of skew (>2) and kurtosis (>4), which can be problematic for latent variable modeling (Kline, 2011). Prior to any deletion of items and set sizes, the alphas for the full working memory span tasks were .86 for operation span, .89 for reading span, and .80 for symmetry span; the overall alpha for the combined tasks (all 42 items) was .93.

Together, the findings in Table 3 serve as a baseline that provides support for our first step in measure reduction: removing items from the lowest set sizes. Table 4 presents the results of the reliability analyses for all of the respective shortened measures. As one would expect mathematically, the alpha coefficients for each task (and across all tasks) will decrease somewhat with the removal of items, and, in our case, the largest difference typically was found between Model 1 (removing the smallest set size) and Model 2 (also removing the third administration). Most alpha reliability coefficients exceeded .70, and many exceeded .75, even in the shortened measures. Reassuringly, Table 4 indicates that the drops in reliability were consistent or slightly less than expectations based on the Spearman-Brown prophecy formula (see Nunnally & Bernstein, 1994), and therefore our measure-shortening strategy had no deleterious effects on the inter-correlation of the remaining items.

Table 2 Measure-reduction models for automated complex span tasks

Model	Number of total items ¹	Description
Model 0	42	<i>Baseline model:</i> Include all items from the operation, reading, and symmetry span tasks included
Model 1	33	<i>First reduced model:</i> Remove items from the smallest set size for operation (set size 3), reading (set size 3), and symmetry span (set size 2)
Model 2	22	<i>Second reduced model:</i> Remove the final administration across all set sizes
Model 3	18	<i>Final reduced model:</i> Remove the largest set size for operation and reading span (both set size 7)

Note. ¹ Items refer to individual administrations of a given set size length

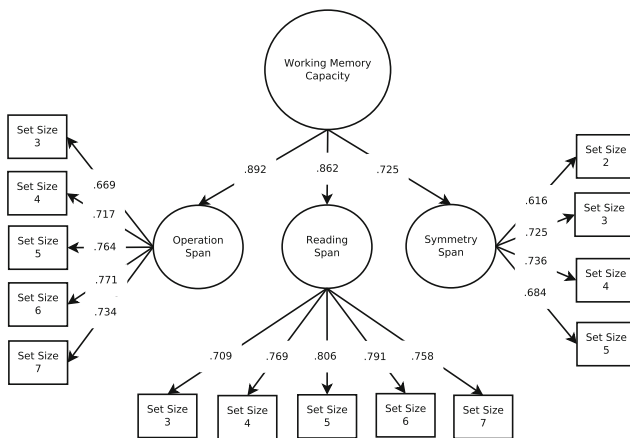
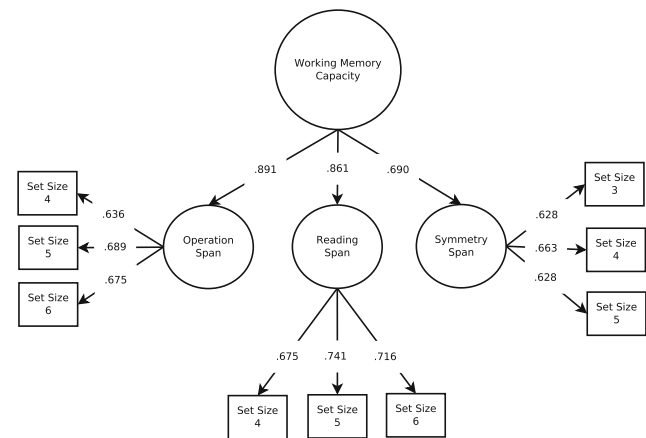


Fig. 1 $N = 2,442$. Confirmatory factor analysis model for the full-length complex working memory span tasks (Model 0) versus shortened counterparts (Model 3). Numbers are standardized factor loadings. Full-length



model: $\chi^2(74) = 501.26, p < .05, CFI = .973, RMSEA = .049 [.045, .053], SRMR = .024$. Shortened model: $\chi^2(24) = 32.52, p = .11, CFI = .999, RMSEA = .012 [.000, .022], SRMR = .010$

Confirmatory factor analyses

Building on these findings with respect to reliability, we tested a two-level hierarchical CFA model as noted previously and as seen in the panels of Fig. 1. Based on this CFA, Table 5 presents the results of all the measure-reduction models. For the baseline model where the measure is not reduced (Model 0), the chi-squared test was statistically significant ($\chi^2(74) = 501.26, p < .05$), indicating poor exact fit for the model. A statistically significant chi-squared test is not unusual for large sample sizes (Kline, 2011); therefore, we also provide commonly reported alternative indices of close model fit, where acceptable close model fit is based on established rules of thumb (i.e., $CFI > .95, RMSEA < .06, SRMR < .08$; Hu & Bentler, 1999). The full measure (Model 0) demonstrated good model fit across all of these alternative indices ($CFI = .973, RMSEA = .049 [.045, .053], SRMR = .024$). The left panel of Fig. 1 presents the model and the standardized factor loadings for Model 0.

For all three measure-reduction models, the CFA models (Models 1–3) also demonstrated a practically significant fit to the data in terms of these alternative model fit indices (see Table 5). In addition, the shortest measure (Model 3) had a non-significant chi-squared value ($\chi^2(24) = 32.522, p = .11$), meaning there was good exact fit as well as close fit. Note that these consecutive measure-reduction models (from Model 0 to Model 3) are non-nested because some indicators (items) are eliminated entirely; this means that we are unable to compare these models statistically using the chi-squared difference test as is often done. However, the Akaike Information Criterion (AIC) values can be compared across non-nested models (Kline, 2011), and as items were removed, the AIC for each consecutive model suggested an improvement in the tradeoff between model fit and model parsimony, meaning the CFA model for the most-reduced measure (Model 3) shows

promise for replicating. (Note, however, that there is no significance test for the difference between AIC values.) Results in Table 5 also suggest a trend of increasing improvements in model fit across models (i.e., decreasing RMSEA and SRMR values, increasing CFI values). Additionally, the general pattern of the factor loadings remained large and consistent across models for the successively reduced measures. Collectively, these findings generally support our measure-shortening strategy. The right panel of Fig. 1 presents the resulting CFA model for the final shortened measure (Model 3).

As a final step, we cross-validated the resulting model for the shortened WMC measures (Model 3) by applying the factor loadings derived from the development sample to an independent cross-validation (holdout) sample of $N = 2,443$ participants (see MacCallum et al., 1994, for this procedure). The statistically significant chi-squared value alone suggests rejecting the cross-validated model ($\chi^2(32) = 48.86, p < .05$), but as noted earlier, chi-squared values are sensitive to the large sample sizes. However, large sample sizes also make indices of close model fit (i.e., conclusions about practical significance) more interpretable, and, here again, those indices were well within accepted standards ($CFI = .997, RMSEA = .015 [.005, .023], SRMR = .023$). These cross-validation results (available on request from the corresponding author) further support the psychometric integrity of a shortened domain-general measure of working memory.

Study 2

Study 2 investigated the reliability and validity of the shortened complex span tasks in an entirely new sample of data. To this end, we administered these tasks to a sample of undergraduate students, along with measures of fluid intelligence (Gf). The specific questions we addressed in this study were

Table 3 Item descriptive statistics, corrected item-total correlations, and reliabilities

Item	Mean	SD	Skew	Kurtosis	IRC
Operation span					
Set Size 3					
Admin. 1	.90	.25	-2.48	5.06	.42
Admin. 2	.92	.22	-2.88	7.53	.45
Admin. 3	.92	.22	-2.89	7.49	.48
Set Size 4					
Admin. 1	.85	.29	-1.88	2.18	.45
Admin. 2	.88	.27	-2.22	3.67	.51
Admin. 3	.90	.26	-2.49	5.05	.54
Set Size 5					
Admin. 1	.79	.33	-1.35	.35	.51
Admin. 2	.82	.30	-1.59	1.25	.52
Admin. 3	.85	.29	-1.82	2.07	.52
Set Size 6					
Admin. 1	.70	.34	-.80	-.72	.49
Admin. 2	.73	.33	-.96	-.37	.53
Admin. 3	.76	.32	-1.11	-.05	.55
Set Size 7					
Admin. 1	.60	.33	-.33	-1.09	.51
Admin. 2	.63	.33	-.42	-1.04	.49
Admin. 3	.64	.32	-.50	-.91	.53
Reading span					
Set Size 3					
Admin. 1	.89	.24	-2.25	4.22	.49
Admin. 2	.88	.26	-2.23	3.95	.49
Admin. 3	.88	.26	-2.16	3.60	.53
Set Size 4					
Admin. 1	.84	.28	-1.73	1.81	.50
Admin. 2	.84	.29	-1.69	1.64	.54
Admin. 3	.84	.29	-1.68	1.59	.59
Set Size 5					
Admin. 1	.78	.32	-1.18	.03	.57
Admin. 2	.77	.32	-1.21	.13	.59
Admin. 3	.77	.33	-1.17	-.02	.59
Set Size 6					
Admin. 1	.69	.33	-.66	-.91	.56
Admin. 2	.67	.34	-.62	-.94	.59
Admin. 3	.67	.35	-.64	-.94	.61
Set Size 7					
Admin. 1	.58	.32	-.21	-1.12	.53
Admin. 2	.56	.33	-.14	-1.17	.54
Admin. 3	.56	.34	-.16	-1.22	.57
Symmetry span					
Set Size 2					
Admin. 1	.91	.21	-2.47	5.58	.33
Admin. 2	.91	.23	-2.50	5.67	.39
Admin. 3	.91	.23	-2.61	6.19	.42

Table 3 (continued)

Item	Mean	SD	Skew	Kurtosis	IRC
Set Size 3					
Admin. 1	.78	.31	-1.20	.22	.44
Admin. 2	.80	.31	-1.30	.45	.47
Admin. 3	.81	.30	-1.39	.71	.49
Set Size 4					
Admin. 1	.64	.35	-.49	-1.12	.47
Admin. 2	.66	.36	-.56	-1.11	.49
Admin. 3	.68	.35	-.66	-.94	.52
Set Size 5					
Admin. 1	.49	.35	.14	-1.28	.45
Admin. 2	.50	.35	.08	-1.28	.47
Admin. 3	.52	.35	.04	-1.32	.48

Note. $N=2,442$. IRC = item-remainder correlation. Coefficient alphas are .86 for operation span, .89 for reading span, .80 for symmetry span, and .93 across all tasks

whether (1) the shortened measures would have acceptable internal consistency reliability, and (2) there would be validity evidence in the form of statistically significant positive correlations between the shortened measures and the *Gf* measures, and a moderate-to-strong positive correlation between latent variables representing WMC and *Gf*, consistent with previous research (e.g., Kane et al., 2005).

Method

Participants

Participants were 185 undergraduate students recruited from the participant pool at Purdue University, a large state university in the Midwest. Participants were compensated with course credit in exchange for their participation. Data from at least one task were missing for 13 participants due to either experimenter error or computer problems, leaving the final sample with $N = 172$ participants with complete data. The final sample had an average age of 19.4 years ($SD = 1.2$) and was 62.2 % female ($N = 107$).

Measures

Working memory capacity Participants completed the shortened versions of the automated operation, reading, and symmetry span tasks. All administration and scoring procedures were identical to those described in Study 1, with the exception that each span task included only those items that were retained for the shortened versions of these tasks. Set sizes for the operation and reading span tasks ranged from 4–6, and set

Table 4 Alpha coefficients for reduced measures

Model	Alpha (short measure)			Predicted alpha (from full measure)			Difference ¹			Mean inter-item correlation					
	Total	Operation	Reading	Total	Operation	Reading	Total	Operation	Reading	Total	Operation	Reading	Symmetry		
Model 0	.93	.86	.89	.80	.89	.86	.86	.76	.01	.01	.01	.24	.29	.35	.26
Model 1	.92	.84	.87	.78	.83	.86	.81	.67	.00	.00	.02	.25	.31	.36	.29
Model 2	.87	.77	.81	.69	.77	.81	.76	.67	.00	.00	.02	.24	.30	.35	.27
Model 3	.84	.71	.76	-	.71	.76	.76	.67	-.01	.00	.01	.23	.29	.35	.27

Note. $N = 2,442$. Predicted alpha is based on applying the Spearman-Brown prophecy formula to the full measure in predicting the reliability for the short measure. Differences are expressed as (alpha – predicted alpha). Model 0 = all items included; Model 1 = smallest set size removed; Model 2 = smallest set size and third administration for all set sizes removed; Model 3 = smallest set size, third administration, and largest set size for operation and reading span removed (not applicable to Symmetry Span). ¹ Difference is based on three-decimal precision and thus may be .01 different from the observed difference between alphas

sizes for the symmetry span task ranged from 3–5; all participants completed two trials at each set size (six items per task).

Paper folding A computerized version of Part 1 of the Paper Folding test (Ekstrom, French, Harman, & Dermen, 1976) was administered. A figure representing a square piece of paper is presented on the left side of the screen, with markings that indicate the “paper” has been folded and then a hole was punched through the paper. Participants were asked to select the answer choice that best represented what that piece of paper would look like if it were completely unfolded. Participants were given a maximum of 5 min to complete the ten items.

Matrix reasoning The current study implemented a computerized version of the Raven’s Advanced Progressive Matrices (e.g., Raven, Raven, & Court, 1998) as a measure of abstract reasoning (*Gf*). The full Raven’s consists of 36 matrix-reasoning items presented in ascending order of difficulty. For each item, participants were presented with a 3 × 3 matrix of abstract shapes, where the bottom-right panel of the matrix is omitted. Participants are asked to select among a number of response alternatives the figure that completes the overall pattern of the matrix. A participant’s total number of correct responses determines their scores on the test. For the present study, we administered only the 18 odd-numbered items from the Raven’s item pool with a maximum time limit of 10 min, consistent with past research (e.g., Kane et al., 2004).

Procedure

Participants were tested in laboratory rooms with up to two other participants. Participants first provided demographic information, and then completed the following measures in order: operation span, symmetry span, reading span, Paper Folding, and Raven’s Advanced Progressive Matrices.²

Results

Table 6 displays descriptive statistics for the shortened span measures and *Gf* measures, along with correlations and reliability estimates. Coefficient alpha was .71 for operation span,

² We wish to report that other data were also collected from a subset of participants (i.e., self-reported ACT/SAT, GPA, hours worked), and other tasks were administered (i.e., antisaccade, arrow flankers, change detection, n-back, and reading comprehension). Note that (1) the short-form automated span tasks were the first three tasks administered in the session; and (2) data from some of the other administered measures was not available for all participants – including only subjects who had complete data would have resulted in a much smaller sample size than what is currently reported.

Table 5 Confirmatory factor analysis models: fit estimates and factor loadings

Model	χ^2 (df)	CFI	RMSEA	SRMR	AIC	Operation		Reading		Symmetry	
						Range (λ)	Mean (λ)	Range (λ)	Mean (λ)	Range (λ)	Mean (λ)
Model 0	501.26 (74)	.97	.049 [.045, .053]	.024	53731.6	.67 - .77	.73	.71 - .81	.77	.62 - .74	.69
Model 1	232.88 (41)	.98	.044 [.038, .049]	.018	46417.6	.70 - .78	.75	.75 - .80	.78	.70 - .75	.72
Model 2	105.92 (41)	.99	.025 [.020, .031]	.014	33576.0	.61 - .69	.67	.66 - .73	.71	.63 - .66	.64
Model 3	32.52 (24)	> .99	.012 [.000, .022]	.010	28063.8	.64 - .69	.67	.68 - .74	.71	.63 - .66	.64

Note. $N=2,442$. Model 0 = all items included; Model 1 = smallest set size removed; Model 2 = smallest set size and third administration for all set sizes removed; Model 3 = smallest set size, third administration, and largest set size for operation and reading span removed. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; AIC = Akaike information criterion. Ranges and means of standardized factor loadings are presented for parcels (groups) of set sizes that load on each of the three corresponding span task factors

.54 for reading span, and .59 for symmetry span, and the coefficient alpha for a composite variable (the total number correct across all three span tasks) was .76. The span measures correlated moderately and positively with each other, and with the *Gf* measures. Although our interpretation of the shortened span measures focuses on latent-variable analyses (see below), we note that the correlations among the span tasks and with the *Gf* measures in the current study are generally similar to previous studies using the full-length span tasks with samples composed entirely of undergraduate students (e.g., Brewer & Unsworth, 2012; Jaeggi et al., 2010; Shelton, Elliott, Hill, Calamia, & Gouvier, 2009; Unsworth, Brewer, & Spillers, 2009; Unsworth & Spillers, 2010).

Confirmatory factor analysis

In order to determine whether the reliability of the shortened span tasks is supported in a new sample, we first conducted a CFA with the shortened span tasks that is the same as the CFA presented in the final model of Study 1 (Model 3; see Table 5). This CFA model demonstrated excellent fit to the data in terms of both exact fit and approximate fit indices (χ^2 (24) = 30.43, $p > .05$, CFI = .978, RMSEA = .038 [.000, .075], SRMR = .044), thus supporting the reliability of the shortened span tasks.

We next performed a CFA as a formal test of validity for the shortened span tasks and the *Gf* measures. Here we specified a

model that correlates two latent factors: WMC, with the three shortened span measures as indicators, and *Gf*, with the two measures of reasoning ability as indicators. There are two important things to note about this model: First, because Study 1 and Study 2 demonstrate support for the reliability of the shortened WMC measures based on CFAs of the *items*, we now use the WMC *measures* themselves in CFA as indicators of the WMC latent factor. Likewise, two reliable measures of *Gf* are used as indicators. Second, the two factor loadings for *Gf* were constrained to be equal, because otherwise the measurement model for *Gf* would be “borrowing” information from the span measures, thus biasing the estimates of the factor loadings and latent correlation between *Gf* and working memory (see Kline, 2011). The standardized solution presented in Fig. 2 will have unequal standardized factor loadings, however, because the error variance estimates going into the standardization were freely estimated and not constrained to equality.

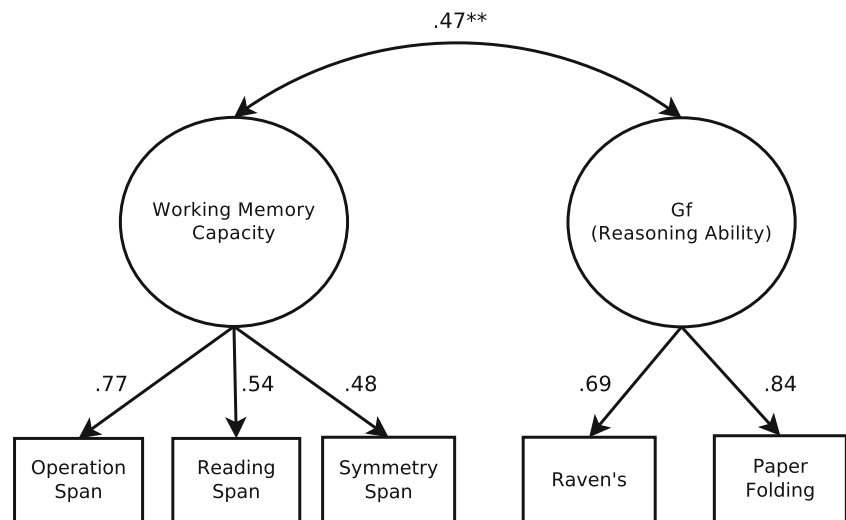
The CFA model demonstrated excellent fit to the data (χ^2 (5) = 3.33, $p > .05$, CFI = >.999, RMSEA = .000 [.000, .082], SRMR = .021). Moreover, each span measure had a strong positive loading on the WMC factor, indicating high factor-based reliability, and the latent WMC factor correlated moderately and positively with the latent *Gf* factor ($r = .47$, $p < .001$; bias-corrected bootstrapped 95 % CI = [.30, .66]), indicating convergent validity between ability constructs (note that this estimate is very close to the meta-analytic correlation

Table 6 Descriptive statistics and correlations for working memory span tasks and ability measures

Task	Mean	SD	Skew	Kurtosis	1.	2.	3.	4.	5.
1. Operation span	23.98	5.75	-1.07	.59	<i>.71</i>	.67	.51	.36	.36
2. Reading span	23.25	4.78	-.42	-.71	.42	<i>.54</i>	.35	.31	.26
3. Symmetry span	15.91	4.33	-.55	.33	.33	.20	<i>.59</i>	.35	.25
4. Paper folding	6.69	2.19	-.50	-.34	.24	.18	.21	<i>.62</i>	.87
5. RAPM	10.08	2.73	.06	-.41	.25	.16	.11	.57	<i>.69</i>

Note. $N = 172$. RAPM = Raven’s Advanced Progressive Matrices. Alpha reliabilities are reported in italics on the main diagonal. Correlations in bold are statistically significant ($p < .05$). Correlations corrected for attenuation due to unreliability are reported above the diagonal

Fig. 2 $N = 172$. Confirmatory factor analysis model with latent variables representing working memory capacity and fluid intelligence (*Gf*). Numbers are standardized factor loadings and the latent correlation. $\chi^2(5) = 3.33$, $p > .05$, CFI = $> .999$, RMSEA = $.000$ [$.000$, $.082$], SRMR = $.021$. $**p < .001$



of $r = .50$ between WMC and Raven's obtained by Ackerman, Beier, & Boyle, 2005).

Because of these encouraging psychometric results, it is worth examining the practical benefit of the short measure in terms of testing time that is conserved. Based on our current and past experiences with these measures, the time to read instructions and practice all of the span tasks (regardless of test form) is approximately 15 min total, so adding this time to the testing times reported in Table 7 means that it took 39.3 min to complete the long form (based on Study 1 data) and 24.7 min on average to complete the short form (based on Study 2 data), a saving of 14.6 min (37.1 %). In addition, when the test is administered in groups (e.g., in a computer lab), the slowest examinee is essentially what determines the length of the testing session. Assuming that the slowest person on average takes 20 min to read instructions (instead of 15) and is at the 95th percentile of the total distribution of times, then this suggests that group sessions will last 58.0 min for the long form and 34.6 min for the short form, a saving of 23.4 min (40.3 %). Although these are rough estimates, they are

reasonable and large enough to suggest that significant administration time will be saved in whatever setting the short form of the complex span working memory tasks is administered.

General discussion

Scores on a variety of established measures of WMC have long been known to correlate highly with one another and demonstrate theoretically appropriate patterns of convergent and discriminant validity, and, in general, cognitive ability measures have long been known to predict a variety of important life outcomes in academic, employment, and personal domains (Gottfredson, 1997). However, implementation of WMC measures—in particular, complex span measures—requires a great deal of time on the part of both administrators and participants; even in the computerized versions of these tasks, roughly 20 min *per task* is a typical administration time

Table 7 Time and time savings for the long versus short working memory span tasks (in min)

Span task	Long form		Short form		Mean time saved
	Mean (SD)	Upper 95 % ile	Mean (SD)	Upper 95 % ile	
Operation	9.8 (3.8)	16.9	3.5 (.7)	4.8	6.3
Reading	10.3 (3.6)	17.3	3.9 (.8)	5.3	6.4
Symmetry	4.2 (1.4)	7.0	2.3 (.5)	3.2	1.9
Total	24.3 (7.0) ^a	38.0 ^a	9.7 (1.5)	13.0	14.6

Note. Long form: $N = 5,003$ for Operation span, $N = 4,342$ for Reading span, $N = 4,993$ for Reading span; Short form: $N = 172$. Times above are for taking the measures themselves; they do not include the fixed time costs of reading instructions and practicing the tasks that, conservatively, adds about 15 min to the total time, regardless of form. ^a These estimates required the reasonable assumption that the correlation between finish times is similar to that for the short form and that the distribution of total times (summing across all item response latencies) is normal

(Unsworth et al., 2005). In some settings, then, it is likely not sensible or even possible to administer several different complex working memory span tasks, given limited testing time and the likely research desire to collect additional data (e.g., experimental data, data on other constructs). Two of the current authors themselves were personally motivated to develop this shortened WMC measure because of testing constraints imposed in recent collaborative work (Barch et al., 2009; Hambrick et al., 2011).

Study 1 took a principled psychometric and conceptual approach to developing a short measure of WMC that samples items across three existing computerized complex span tasks. Our model-based CFA approach to reducing a measure of WMC was based on a large development sample used to estimate the model parameters, then the model parameters were applied to data from a similarly large hold-out cross-validation sample, yielding supportive model fit results. Study 2 produced evidence for the reliability and validity of the shortened span measures in a new sample: the CFA of the shortened working memory measure provided additional support for its reliability in an independent sample. Furthermore, the correlation between WMC and *Gf* latent variables was .47, within the range expected based on previous research with the full-length span tasks (e.g., Redick, Unsworth, Kelly, & Engle, 2012; Unsworth & Spillers, 2010). This psychometric support is required before considering the practical benefits of reduced test administration time, where we found a savings of anywhere from 15–20 min on the test itself, which shaves about 40 % off of the total administration time.

To summarize, the current study developed and executed a principled procedure for developing and psychometrically modeling and testing a shortened measure of overall WMC. We believe the short measure offers numerous practical benefits to future research by allowing working memory to be assessed quickly, to allow researchers to measure other constructs within the natural limits of testing time and the natural limits on test-taker's patience. Note that the shortened WMC is most obviously used in individual-differences studies, but it can also be incorporated into experimental designs that take individual differences into account (e.g., for use as covariates, a priori in matched designs, or post hoc in propensity-score matching).

Future research The current research might inspire a number of additional avenues for future research. First, although a wide range of set sizes was investigated, future research could examine a wider range of set sizes (e.g., somewhat larger set sizes for symmetry span). Naturally, any changes to the measures would require additional model testing, and the current research could serve as a general guide for doing so. Future research in this direction might also include other types of span tasks (e.g., counting span; Case, Kurland, & Goldberg, 1982; Engle et al., 1999) and might incorporate samples displaying a wider or different range of working memory

capacity (e.g., age-diverse community samples; samples with span-specific expertise). The short working memory measure developed here should also demonstrate expected patterns of convergent and discriminant validity with other cognitive constructs such as inhibition (e.g., Hasher & Zacks, 1979; Miyake et al., 2000) and proactive interference (Lustig, May, & Hasher, 2001), and motivational constructs such as resource depletion (Muraven & Baumeister, 2000; Schmeichel, 2007) and cognitive fatigue (Ackerman & Kanfer, 2009; Ackerman, 2011). Future research might examine specific span tasks in this nomological network as well, to help researchers with limited testing time make a decision between administering a domain-specific complex span task in full, versus administering the current short measure of overall WMC, which samples heterogeneously across span tasks, capitalizes on the general factor, and keeps testing time at an efficient level.

Second, although our measure-shortening efforts were in the service of developing a domain-general measure of working memory with heterogeneous content across span measures, there is certainly additional utility that might be gained in creating a shorter measure within a specific span task or domain of working memory. For example, it may be most appropriate to measure specific span tasks or domains at the exclusion of other domains within certain clinical populations or gifted-student populations to assess their level of anticipated deficits or strengths. Shortened span measures for these populations might be psychometrically tailored in a different manner. Likewise, if a shortened measure of general WMC were ever used to make decisions about individuals, then it might need to be tailored to ensure high conditional reliability around the relevant cut-scores.

The final consideration is a forward-looking methodological step (and not necessarily a requirement) with regard to our cross-validation strategy for the shortened measure. We constrained the factor loadings of the fully reduced model in our cross-validation sample to equal to those for the development sample. Our findings indicated good model fit in the cross-validation sample; however, other models and approaches to cross-validation are possible: more restrictive models might fix the latent variances and possibly the error variances as well (see MacCallum et al., 1994), or another approach to cross-validation might be a Bayesian CFA, where small cross-loadings and residual correlations can be reasonable to specify and simple to estimate (Muthén & Asparouhov, 2012), and for cross-validation purposes (essentially), the estimated probability distribution of parameters derived from a development sample could provide prior information that informs subsequently estimated distribution in the cross-validation data set. Also, future research providing information on convergent, discriminant, and criterion-related validity might suggest additional models for reducing the measure, especially when the goal is to estimate relative patterns of relationship between variables (vs. predicting individual

scores), or when jointly considering the measurement and value of other constructs given limited testing time.

Author note Samuel T. McAbee is currently a visiting assistant professor at Illinois Institute of Technology. Our short working memory measure is available online at <http://englelab.gatech.edu/tasks.html>.

References

- Ackerman, P. L. (2011). (Ed.) *Cognitive fatigue: Multidisciplinary perspectives on current research and future applications*. Washington, DC: American Psychological Association.
- Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*, 30–60.
- Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*, 163–181.
- Arthur, W., & Day, D. (1994). Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement*, *54*, 395–403.
- Barch, D. M., Berman, M. G., Engle, R., Jones, J. H., Jonides, J., MacDonald, A. III ... Sponheim, S. R. (2009). CNTRICS final task selection: Working memory. *Schizophrenia Bulletin*, *35*, 136–152.
- Beier, M. E., & Ackerman, P. L. (2005). Working memory and intelligence: Different constructs. Reply to Oberauer et al. (2005) and Kane et al. (2005). *Psychological Bulletin*, *131*, 72–75.
- Bors, D. A., & Stokes, T. L. (1998). Raven's advanced progressive matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*, *58*, 382–398.
- Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language*, *66*, 407–415.
- 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.
- Colom, R., Escorial, S., Shih, P. C., & Privado, J. (2007). Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents.
- 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., Elliott, E. M., Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. 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.
- Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422–433.
- Deary, I. J., Weiss, A., & Batty, G. D. (2010). Intelligence and personality as predictors of illness and death: How researchers in differential psychology and disease epidemiology are collaborating to understand and address health inequalities. *Psychological Science*, *11*, 53–79.
- Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.
- Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19–23.
- Engle, R. W., Cantor, J., & Carullo, J. J. (1992). Individual differences in working memory and comprehension: A test of four hypotheses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 972–992.
- 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: 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.
- Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2014). Shortened complex span tasks can reliably measure working memory capacity. *Memory and Cognition*.
- Friedman, N. P., & Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory & Language*, *51*, 136–158.
- Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*, 79–132.
- Hambrick, D. Z., Engle, R. W., & Kane, M. J. (2004). The role of working memory in higher level cognition: Domain-specific vs. domain-general perspectives. In R. J. Sternberg & J. Pretz (Eds.), *Intelligence and cognition*. New York: Cambridge University Press.
- Hambrick, D. Z., Rench, T. A., Potoski, E. M., Darowski, E. S., Roland, D., Bearden, R. M. ... Brou, R. (2011). The relationship between the ASVAB and multitasking in Navy sailors: A process-specific approach. *Military Psychology*, *23*, 365–380.
- Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, *108*, 356–388.
- Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist*, *17*, 475–483.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.
- Hull, R., Martin, R. C., Beier, M. E., Lane, D., & Hamilton, A. C. (2008). Executive function in older adults: A structural equation modeling approach. *Neuropsychology*, *22*, 508–522.
- Jaeggi, S. M., Studer-Luethi, B., Buschkuhl, M., Su, Y.-F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning - Implications for training and transfer. *Intelligence*, *38*, 625–635.
- Jeyakumar, S. L. E., Warriner, E. M., Raval, V. V., & Ahmad, S. A. (2004). Balancing the need for reliability and time efficiency: Short forms of the Weschler Adult Intelligence Scale-III. *Educational and Psychological Measurement*, *64*, 71–87.
- Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The Big Five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*, 621–652.
- 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: Oxford University Press.
- Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 66–71.
- Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity:

- A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*, 189–217.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, *9*, 151–173.
- Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When “good” indicators are bad and when “bad” indicators are good. *Psychological Methods*, *4*, 192–211.
- Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General*, *130*, 199–207.
- MacCallum, R. C., Roznowski, M., Mar, C. M., & Reith, J. V. (1994). Alternative strategies for cross-validation of covariance structure models. *Multivariate Behavioral Research*, *29*, 1–32.
- Miller, H. R., Streiner, D. L., & Goldberg, J. O. (1996). Short, shorter, shortest: The efficacy of WAIS-R short forms with mixed psychiatric patients. *Assessment*, *2*, 165–169.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100.
- Muraven, M., & Baumeister, R. G. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, *126*, 247–259.
- Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335.
- Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working memory and intelligence—Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 61–65.
- Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity—Facets of a cognitive ability construct. *Personality and Individual Differences*, *29*, 1017–1045.
- Raven, J. C., Raven, J., & Court, J. H. (1998). *A manual for Raven's progressive matrices and vocabulary scales*. London: H. K. Lewis.
- Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012a). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*, 164–171.
- Redick, T. S., Unsworth, N., Kelly, A. J., & Engle, R. W. (2012b). Faster, smarter? Working memory capacity and perceptual speed in relation to fluid intelligence. *Journal of Cognitive Psychology*, *24*, 844–854.
- Schmeichel, B. J. (2007). Attention control, memory updating, and emotion regulation temporarily reduce the capacity for executive control. *Journal of Experimental Psychology: General*, *136*, 241–255.
- Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*, 162–173.
- Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General*, *125*, 4–27.
- Shelton, J. T., Elliott, E. M., Hill, B. D., Calamia, M. R., & Gouvier, W. D. (2009). A comparison of laboratory and clinical working memory tests and their prediction of fluid intelligence. *Intelligence*, *37*, 283–293.
- Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*, 167–194.
- Unsworth, N., Brewer, G. A., & Spillers, G. J. (2009a). There's more to the working memory capacity–fluid intelligence relationship than just secondary memory. *Psychonomic Bulletin & Review*, *16*, 931–937.
- 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.
- Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009b). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory*, *17*, 635–654.
- Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention, memory, or both? A direct test of the dual-component model. *Journal of Memory and Language*, *62*, 392–406.
- Was, C. A., Rawson, K. A., Bailey, H., & Dunlosky, J. (2011). Content-embedded tasks beat complex span for predicting comprehension. *Behavior Research Methods*, *43*, 910–915.