Refereed Proceedings of Doing Psychology:
Manawatū Doctoral Research Symposium Volume 2
19-26

Into the Void: The Gap Between N-Back and Complex Span Tasks Suggests Inadequacies in Current Models of Working Memory

Kathryn Campbell, Stephen Hill & John Podd
kacey.a.campbell@gmail.com, S.R.Hill@massey.ac.nz, J.V.Podd@massey.ac.nz
School of Psychology, Massey University, Palmerston North, New Zealand

Abstract

The tasks used to assess working memory are a highly contentious issue in cognitive psychology. Previous research has found a weak relationship between two key types of working memory tasks: N-Back and Complex Span. This is commonly interpreted as evidence that one or both tasks possess poor construct validity. However, this finding may be a result of assessing different modalities of working memory. The current pilot study aimed to clarify the differences between the two tasks by assessing performance on each within the same modality. A spatial and verbal version of each task was used. Although, theoretically, these tasks assess the same construct, the pilot data revealed low correlations between them. This suggests that the current models of working memory may be inadequate, or that unidentified differences between the tasks may be influencing the results. Due to their widespread use and applications, it is important to better understand models of working memory and develop improved tasks.

Keywords: Working memory, N-Back, Complex span, Spatial working memory, Verbal working memory

Background Research

What is Working Memory?

Working Memory (WM) is a short-term store for maintaining and processing information (Baddeley & Hitch, 1974). WM is distinguished from Short-Term Memory (STM) due to its active component. STM involves simple rehearsal and is sometimes labelled as a 'passive store' (Swanson, 1994). On the other hand, WM is an active store involving rehearsal and processing of stimuli. For instance, STM is used to remember the digits of a phone number while WM would be used to maintain the same numbers while also trying to decide what to say when the call connects. A distinction between STM and WM is supported by factor analysis suggesting that the two constructs, although related, are distinct (e.g., Kail & Hall, 2001; Swanson, 1994). Furthermore, WM and STM differentially predict performance in other cognitive tasks such as word decoding and measures of fluid intelligence (Kail & Hall, 2001; Swanson, 1994).

There are numerous models of WM (See Miyake & Shah, 1999 for a review). The dominant model was proposed by Baddeley and Hitch (1974). In this model WM is composed of three key subsystems: A central executive, a phonological loop and a visuo-spatial sketchpad. The central executive essentially oversees all WM processes. The phonological loop and the visuo-spatial sketchpad are systems to rehearse and temporarily store modality-specific information. However, not all models agree and there is debate regarding the very existence of the central executive (Parkin, 1998) and the existence of separate resource pools for different modalities (Miyake & Shah, 1999).

WM is related to numerous and varied factors such as intelligence (Kane, Hambrick, & Conway, 2005), reading comprehension (Daneman & Carpenter, 1980), and ability to deal with stressful events (Klein & Boals, 2001). Furthermore, Working Memory Capacity (WMC) appears to be deficient in mental health disorders including schizophrenia and depression (Barch, Sheline, Csernansky, & Snyder, 2003; Rose & Ebmeier, 2006). The apparent usefulness of the WM construct highlights the need for valid and reliable measures to assess WM capacity. However, there is debate regarding which tasks should be used and evidence that various WM tasks are not measuring the same thing.

How is Working Memory Measured?

Two commonly used WM measures are N-back tasks and Complex Span Tasks (CST's). A CST involves remembering a string of items while completing a secondary, sometimes unrelated, task. For example, Shah and Miyake (1996) designed a spatial CST (Figure 1) in which participants were required to indicate whether a string of letters were ‘normal’ or ‘mirror-imaged’ while also remembering the degree of rotation for every letter. The longest string of letters in which the participant could correctly recall all letter rotations was indicative of their WM capacity.

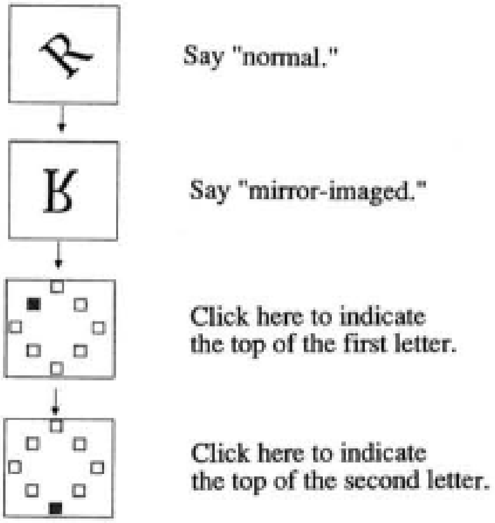
Say "normal."
Say "mirror-imaged."
Click here to indicate the top of the first letter.
Click here to indicate the top of the second letter.

Figure 1. Example of a spatial working memory task designed by Shah and Miyake (1996). Participants completing this task must decide if a series of letters are ‘normal’ or ‘mirror-imaged’. The participants must then click on a square to indicate the direction that the top of the letters were pointing.

N-Back tasks involve a continuous string of items in which the participants’ task is to identify if the currently displayed item is the same as the one displayed a set number of items ago (Figure 2).

Comparison of N-Back and Complex Span

Proponents of the N-Back task commonly justify its use claiming that it possesses face validity (Jaeggi, Buschkuehl, Perrig, & Meier, 2010; Kane, Conway, Miura, & Colflesh, 2007). During N-Back tasks participants must rehearse/store a set of stimuli while continuously updating this set and responding to all of the presented stimuli. This set of tasks appears to satisfy the two key components of WM theory: simultaneous storage and processing.

However, there have been few empirical evaluations of N-Back tasks and the results that do exist provide mixed findings (Jaeggi et al., 2010). N-Back tasks have been found to correlate more strongly with simple, STM tasks than CST’s (Jaeggi et al., 2010; Roberts & Gibson, 2002) and account for the same variance in language comprehension performance as STM tasks (Kwong See & Ryan, 1995). This suggests that N-Back tasks may be assessing STM capacity rather than WM. Jaeggi et al. (2010) also provided evidence that N-Back tasks suffer from poor reliability. On the other hand, performance in N-Back tasks correlates well with performance in domains thought to be influenced by WM capacity such as intelligence (e.g., Aronen, Vuontela, Steenari, Salmi, & Carlson, 2005; Gevins & Smith, 2000). Additionally, N-Back tasks may elicit similar patterns of brain activity as other tasks used to assess attention control in WM tasks (Gray, Chabris, & Braver, 2003). Given such mixed evidence regarding the validity of the N-Back task, the results of studies relying solely on it to assess WM are called into question.

In comparison, CST’s have been extensively validated. CST’s also have face validity due to the dual storage and processing requirements (Kane et al., 2007). CST’s correlate more strongly with memory tasks involving manipulation of the stimuli than simple STM tasks (e.g., Engle, Tuholski, Laughlin, & Conway, 1999) and predict abilities on domains thought to require WM performance but not domains that do not (Engle

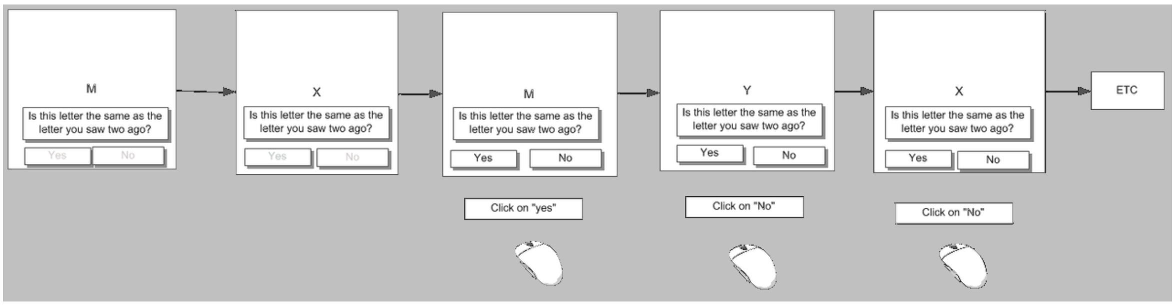
M
Is this letter the same as the letter you saw two ago?
Yes
No
X
Is this letter the same as the letter you saw two ago?
Yes
No
M
Is this letter the same as the letter you saw two ago?
Yes
No
Click on "yes"
Y
Is this letter the same as the letter you saw two ago?
Yes
No
Click on "No"
X
Is this letter the same as the letter you saw two ago?
Yes
No
Click on "No"
ETC

Figure 2. In a verbal N-Back task, the participant is shown a continuous string of letters. For every letter displayed the participant must state whether the letter displayed matches the one displayed a set number of items ago. This figure displays a 2-back version of the task.

& Kane, 2004). These findings suggest that CST's possess good construct validity. Additionally, a review focussing on automated versions of several CST's found good test-retest reliability, construct and criterion-related validity, convergent validity and internal consistency (Redick et al., 2012).

While there is evidence that both types of WM tasks possess validity, the relationship between them is substantially weaker than would be expected (*r* ~.2) if they were both assessing the same construct (Jaeggi et al., 2010; Kane et al., 2007; Oberauer, 2005; Roberts & Gibson, 2002). Furthermore, although both tasks explain variance in performance on tasks requiring WM, they may do so independently (Jaeggi et al., 2010; Kane et al., 2007). That suggests the two task types may be measuring different constructs.

Why, if both tasks possess face and construct validity, do they not show a higher correlation? It is possible that the relationship between the two tasks is being masked by the N-Back task's poor reliability (Jaeggi et al., 2010). However, it is likely that the low correlation results from uncontrolled, task-specific variance.

One of the most striking problems is that modality or content-specific task variance has rarely been taken into account. Many dominant models of WM propose multiple pools of resources for different modalities (Miyake & Shah, 1999). Evidence supports these theories as WM performance has been found to be dependent upon the modality which the stimuli in the task belong to (Perlow, Moore, Kyle, & Killen, 1999). Modality-specific brain activity has also been observed (Jaeggi et al., 2010; Smith & Jonides, 1997). Previous research comparing CST's and N-Back tasks has rarely controlled for modality-specific effects. For instance, Kane et al. (2007) conducted research comparing Operation Span (a numerical and language-based CST) with a letter-based N-Back task. Such a comparison may have been confounded by individuals' performance in the non-shared numerical component of the CST. In an effort to resolve this problem, two methods have been used. Firstly, some research has used data averaged across multiple versions of each type of task covering a variety of modalities. For example, Shamosh et al. (2008) found higher correlations than those comparing performance on a single version of each task (e.g., *r*=.55). However, averaging data across multiple tasks limits the potential for comparison of different modalities and can only reduce, not eliminate, the impact of task/modality-specific variance.

A second method used to assess the impact of modality employed a latent variable approach (e.g., Schmiedek, Hildebrandt, Lövdén, Wilhelm, & Lindenberger, 2009). This method is thought to minimise the impact of error and task-specific variance allowing for an unbiased comparison of tasks. Using three versions of each type of task, assessing multiple modalities, Schmiedek et al. found evidence of a single latent factor common to both task types. This factor was thought to represent an aspect of WM such as controlled attention. Schmiedek et al. interpreted this result as evidence that the N-Back tasks assess essentially the same construct as CST and are therefore equally as valid.

Despite attempts to clarify the cause of the poor relationship between N-Back and CST's, no conclusion has been widely accepted and further research is required. If no task-specific factors can be identified as the source of the low correlations between the tasks, this may indicate that the theories of WM upon which these tasks are designed and validated are incomplete.

Pilot Research

Overview

As part of a larger research project investigating the role of brain activity in WM deficits in depression, two sets of WM tasks were piloted. These tasks were devised in order to compare spatial and verbal WM function. After reviewing the literature regarding the best WM task to use in this context, it was decided to develop two tasks (CST and N-Back) for each type of WM (Verbal and Spatial). This approach ensured that all aspects of WM were captured and also provided material for the on-going debate regarding the use of these tasks in WM research. To provide an accurate comparison between verbal and spatial WM it was important that the tasks were equivalent in difficulty and design.

Method

Participants

Data from 16 participants were used for analysis. The participants for the pilot study were students at Massey University. No other demographic data were recorded.

Materials and Procedure

Four WM tasks were required for this study. Two tasks were used to assess Spatial WM and two tasks were used to assess Verbal WM. For each modality of WM both an N-Back and a CST was required. Due to the requirement of matched tasks, both N-Back tasks were designed from scratch for the current research.

For the N-Back tasks the participants were required to judge whether the stimulus presented on the screen matched the stimulus displayed a set number of items ago - designated ‘N’. Each N-Back task consisted of four blocks of trials. Two blocks presented the 2-back condition and two blocks used the 3-back condition. Each block contained 40 stimulus presentations. Eight stimuli were presented in a counter-balanced order in a continuous string of 40 stimuli. For the verbal N-Back task (Figure 2) the stimuli were eight phonologically distinct letters: M, H, K, Q, X, R, F, and B. For the spatial N-Back task (Figure 3) the stimuli were squares located in one of eight equidistant locations surrounding a fixation cross in the centre of the screen. The squares were located at 0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315°. Each block began with a fixation cross in the centre of the screen for 500 ms. Each stimulus was then displayed on the screen for 3000 ms. If participants did not respond during this time they were recorded as incorrect and the next stimulus was displayed. The inter-stimulus interval was 500 ms. The participants responded by using their mouse to select ‘yes’ or ‘no’ in response to the question: “Is this (stimulus) the same as the one displayed n-back ago?” For each task the participants were provided with a set of instructions and were then given two short practice blocks. During the practice blocks participants were provided with immediate audio feedback for their performance. A ‘ding’ signified a correct answer while a ‘dong’ sounded if the participant was incorrect or failed to respond within the time limit. This helped to ensure participants had understood the instructions.

A pre-existing verbal CST, known as the Automated Reading Span Task, was acquired for this research (Unsworth, Heitz, Schrock, & Engle, 2005). In this task the participants were shown a sentence and were asked to decide if it made sense (50% did) or not. Those that did not contained a single word that did not make sense in context. After verifying the sentence, the participants were shown a letter. The participants were asked to complete this pair of tasks a number of times before recalling the letters in the correct order. The number of verification/letter pairs within a block ranged from two to five. Each block length was repeated three times during the experiment resulting in 12 blocks of trials.

Finally, the spatial CST was designed using Shah and Miyake’s (1996) task as a guide (Figure 1). In this task the participants were asked to indicate if a series of rotated letters were ‘normal’ or ‘mirror-imaged’. Following a series of these decisions, the participants were asked to recall the angle of rotation for each of the letters from the previous string by using arrows (See Figure 4). Three phonologically distinct letters were used as the stimuli: F, P and R. The stimuli were rotated around a fixed point. With 0° representing an upright letter, the stimuli were displayed at seven different orientations: 45°, 90°, 135°, 180°, 225°, 270°, and 315°. Each stimulus could be presented in the ‘normal’ position or ‘mirror-imaged’. Therefore, there were a total of 42 unique stimuli. Based on findings by Shah and Miyake (1996) it was determined that the maximum string of letters should be 6. Each string size (1, 2, 3, 4, 5, or 6) was presented twice to give each participant two chances to correctly recall the letter orientations within the string resulting in 12 blocks of trials.

For both CST’s, the absolute and total scoring methods were used. An absolute score is

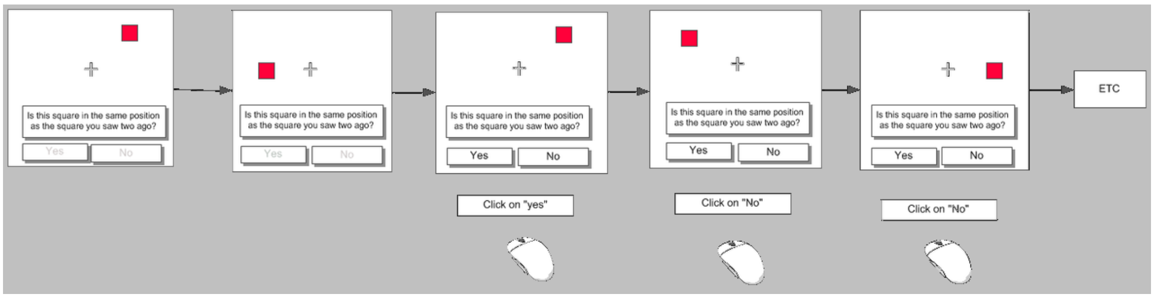
Is this square in the same position as the square you saw two ago?
Yes
No
Is this square in the same position as the square you saw two ago?
Yes
No
Is this square in the same position as the square you saw two ago?
Yes
No
Click on "yes"
Is this square in the same position as the square you saw two ago?
Yes
No
Click on "No"
Is this square in the same position as the square you saw two ago?
Yes
No
Click on "No"
ETC

Figure 3. In this spatial N-Back task the participant must decide if the square is located in the same position as the square two screens previous.

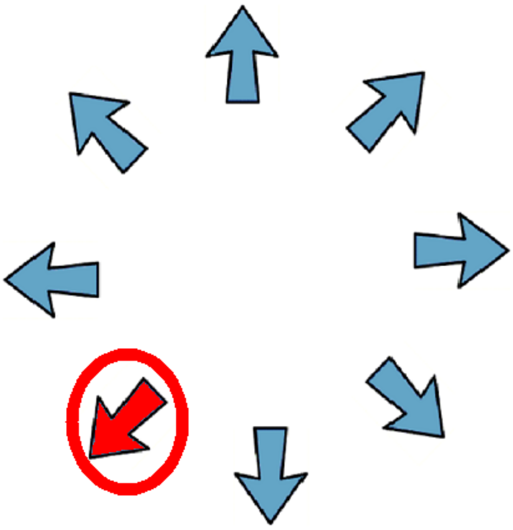

Figure 4. During the recall of rotation phase of the spatial CST the participants use arrows to indicate the direction that the top of the letters were pointing.

determined by adding up *only* the stimuli recalled from perfectly recalled blocks (i.e., blocks in which *all* items are recalled correctly). For example, in a block of five stimuli, if a participant made a single mistake then none of the stimuli within that block would count towards the absolute score. On the other hand, a total score provides the total number of stimuli recalled correctly, regardless of whether the rest of the block was recalled correctly. Research suggests that the total span score may be more reliable and provide a better indicator of individual differences (Conway et al., 2005; Friedman & Miyake, 2005; Redick et al., 2012).

Table 1.
Relationships among Performances in the Four WM Tasks

		Spatial N-Back		Verbal N-Back		Spatial Span		Verbal Span	
		2-Back	3-Back	2-Back	3-Back	Total	Absolute	Total	Absolute
Spatial N-Back	2-Back	___							
	3-Back	.762**	___						
	Mean	.937**	.941**						
Verbal N-Back	2-Back	.279	.294	___					
	3-Back	.335	.407	.844**	___				
	Mean	.318	.361	.965**	.955**				
Spatial Span	Total	.292	.278	.304	.357	___			
	Absolute	.247	.201	.307	.372	.943**	___		
Verbal Span	Total	.174	.334	.103	.350	.147	.026	___	
	Absolute	.138	.420	.095	.385	.268	.134	.884**	___

Results and Discussion

Table 1 displays the relationships among performances in the four WM tasks. The correlations between the two spatial WM tasks (.201-.292), and between the two verbal WM tasks (.095-.385), are considerably lower than would be expected if both the N-Back and CST task both assessed performance on the same construct. However, both tasks possess face validity and appear to fit well with most currently accepted models of WM. Therefore, this result could indicate a gap in the current models of WM. For example, the tasks may assess different sub-components of WM.

Another possibility is that a fundamental difference remains between the two tasks types. For example, Jaeggi et al. (2010) suggested that N-Back tasks rely primarily on recognition memory, while CST's rely primarily on recall. In general, performance in recall tasks appears to be poorer than in recognition tasks, possibly reflecting differences in difficulty and/or underlying processes (Haist, Shimamura, & Squire, 1992). Recognition memory can be thought of as being a dual-process involving both recollection and familiarity, while recall tasks rely solely on recollection due to the absence of cues required to make familiarity judgements (Rugg & Yonelinas, 2003). In CST's the cues available are limited and participants need to rely on accurate recollection to perform well. However, in N-Back tasks participants are provided with cues so may employ both recollection and familiarity processes. The use of different underlying processes may reduce the shared variance observed between CST's and N-Back tasks. While this could indicate that one or both tasks are problematic, it is of concern that current models of WM do not allow for the use of different processes. Models of long-term memory distinguish between implicit and explicit memory processes. Perhaps a similar model of WM with branches for different underlying processes is required.

The correlations between the spatial and verbal versions of the N-Back tasks (.279-.407) were also very low. The spatial and verbal versions of these tasks were identical other than the stimuli used. Therefore, the low correlation can be attributed to modality. Thus, it is important to control for modality or stimulus type when assessing WM. A similar pattern is observed with the CST tasks (.026-.268). This result provides further support for models of WM proposing different pools of resources for different content.

Conclusion

Based on these results, several conclusions can be drawn. Firstly, the low correlation between the N-Back tasks and CST suggests that the tasks may not assess the same construct indicating that researchers should be cautious when interpreting results of studies using these tasks to measure WM. Both task types should be used in future research to enable further investigation of the low correlation.

Secondly, the low correlation between the spatial and verbal variants of both tasks indicates the modality of tasks is important. Stating that an individual suffers from WM deficits should always be further described by elucidating the types of stimuli that led to the observed deficits. Furthermore, the low correlations between spatial and verbal variants of the same task also provide evidence for the multiple resource pool models of WM.

The above findings came from a low N pilot study and need to be replicated with a larger sample. During the pilot study, small changes were made to the instructions for the tasks based on observation and participant feedback. This extra source of variance may have contributed to the low correlations. However, only participants who appeared to have understood the instructions were included in these analyses so these changes should have had a minimal impact.

Future research should further investigate the possibility of different underlying processes in CST's and N-Back tasks with a focus on the use of recall and recognition memory. A WM model that could account for the use of different processes would be beneficial.

In sum, based on current theories of WM, both CST's and N-Back tasks meet the basic requirements of a WM task: simultaneous storage and processing of stimuli. However, given the very low correlations between these tests, it is clear that there is more to WM than these two simple components. The current models of WM need to be reassessed to explain the low correlations. In the meantime, researchers examining WM should use a mixture of WM tasks.

Acknowledgement

I would like to thank my supervisors, Dr. Stephen Hill and Associate Professor John Podd for their help and support with writing this paper and with my research. I would also like to thank Ann Rogerson and Stephanie Denne for organising the 2012 Doing Psychology: Manawatū Doctoral Research Symposium.

Kathryn Campbell is a PhD candidate at Massey University, Palmerston North. She is currently in the data collection phase of her research which she hopes to complete by the end of 2013. Her research interests include asymmetric brain activity in depression, brain asymmetry in self-injurious behaviour, and the link between depression and chronic health conditions.

References

Aronen, E. T., Vuontela, V., Steenari, M. R., Salmi, J., & Carlson, S. (2005). Working memory, psychiatric symptoms, and academic performance at school. *Neurobiology of Learning and Memory, 83*, 33-42.

Baddeley, A.D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation, 8*, 47-89.

Barch, D., Sheline, Y., Csernansky, J., & Snyder, A. (2003). Working memory and prefrontal cortex dysfunction: specificity to schizophrenia compared with major depression. *Biological Psychiatry, 53*, 376-384.

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

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

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. *Psychology of Learning and Motivation: Advances in Research and Theory, 44*, 145-199.

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.

Friedman, N., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods, 37*, 581-590.

Gevins, A., & Smith, M. E. (2000). Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral Cortex, 10*, 829-839.

Gray, J. R., Chabris, C. F., & Braver, T. S.(2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience, 6*, 316-322.

Haist, F., Shimamura, A.P., & Squire, L.R. (1992). On the relationship between recall and recognition memory. *Journal of Experimental Psychology, 18,* 691-702.

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.

Kail, R., & Hall, L.K. (2001). Distinguishing short-term memory from working memory. *Memory & Cognition, 29*, 1-9.

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 question of construct validity. *Journal of Experimental Psychology, Learning, Memory and Cognition, 33*, 615-622.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. (2005). Working memory capacity and fluid intelligence are strongly related constructs: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*, 66-71.

Klein, K., & Boals, A. (2001). The relationship of life event stress and working memory capacity. *Applied Cognitive Psychology, 15*(5), 565-579.

Kwong See, S. T., & Ryan, E. B. (1995). Cognitive mediation of adult age differences in language performance. *Psychology and Aging, 10*, 458-468.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control.* Cambridge: Cambridge University Press.

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.

Parkin, A. (1998). The central executive does not exist. *Journal of the International Neuropsychological Society, 4*, 518-522.

Perlow, R., Moore, D. D. W., Kyle, R., & Killen, T. (1999). Convergent evidence among content-specific versions of working memory tests. *Educational and Psychological Measurement, 59*, 866-877.

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

Roberts, R., & Gibson, E. (2002). Individual differences in sentence memory. *Journal of Psycholinguistic Research, 31*, 573-598.

Rose, E., & Ebmeier, K. (2006). Pattern of impaired working memory during major depression. *Journal of Affective Disorders, 90*, 149-161.Rugg, M.D., & Yonelinas, A.P. (2003). Human recognition memory: a cognitive neuroscience perspective. *Trends in Cognitive Sciences, 7,* 313-319.

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1089-1096.

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*, 4-26.

Shamosh, N. A., DeYoung, C. G., Green, A. E., Reis, D. L., Johnson, M. R., Conway, A. R. A., ... Gray, J. R. (2008). Individual differences in delay discounting: Relation to intelligence, working memory, and anterior prefrontal cortex. *Psychological Science,19*, 904-911.

Smith, E., & Jonides, J. (1997). Working memory: A view from neuroimaging. *Cognitive psychology, 33,* 5-42.

Swanson, H. L. (1994). Short-term memory and working memory: Do both contribute to our understanding of academic achievement in children and adults with learning disabilities? *Journal of Learning Disabilities, 27*, 34-50.

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.

This page intentionally left blank.