Published in final edited form as:

Cogn Affect Behav Neurosci. 2014 March ; 14(1): 106–116. doi:10.3758/s13415-013-0235-x.

Capacity Estimates in Working Memory: Reliability and Interrelationships Among Tasks

Jared X. Van Snellenberg

Department of Psychiatry, Columbia University College of Physicians and Surgeons; Division of Translational Imaging, New York State Psychiatric Institute

Andrew R. A. Conway

Department of Psychology, Princeton University

Julie Spicer

Department of Psychology, Columbia University

Christina Read

Division of Translational Imaging, New York State Psychiatric Institute

Edward E. Smith*

Department of Psychology, Columbia University; Division of Cognitive Neuroscience, New York State Psychiatric Institute

Abstract

The concept of capacity has become increasingly important in discussions of working memory (WM), in so far as most models of WM conceptualize it as a limited-capacity mechanism for maintaining information in an active state, and capacity estimates from at least one type of WM task, complex span tasks, are valid predictors of real-world cognitive performance. However, the term capacity is also often used in the context of a distinct set of WM tasks, change detection tasks, and may or may not refer to the same cognitive capability. We here develop maximum-likelihood models of capacity from each of these tasks, in addition to a third WM task with heavy demands on cognitive control, the self-ordered WM task (SOT), and show that capacity estimates from change detection and complex span tasks are not correlated with each other, although capacity from change detection tasks do correlate with those from the SOT. Furthermore, exploratory factor analysis confirmed that performance on the SOT and change detection load on the same factor, with performance on our complex span task loading on its own factor. These findings suggest that at least two distinct cognitive capabilities underlie the concept of WM capacity as applied to each of these three tasks.

Introduction

One of the hallmarks of Edward E. Smith's approach to scientific problems was an eagerness to use all available tools to constrain hypotheses and distinguish between alternative explanations of data. In our many discussions of human working memory (WM), he was always struck with the elegance and utility of the simple mathematical model of WM capacity used in the literature on change detection tasks (Cowan, 2000; Pashler, 1988), and as we progressed in our research into WM deficits in patients with schizophrenia at the end of his career (Smith & Van Snellenberg, 2011; Van Snellenberg et al., 2013; Van

*Edward Smith (deceased August 17, 2012) was heavily involved throughout the preliminary and intermediate phases of the study reported in this manuscript. He saw early versions of the analyses presented here, but passed away before the maximum likelihood models were fully developed. However, he had seen and approved of earlier analyses that broadly parallel those presented here.

Snellenberg, Wager, Abi-Dargham, Urban, & Smith, 2010), he was eager to develop a similar capacity model for the task we were using to probe these deficits, the self-ordered working memory task (SOT). A critical question was whether the ability to hold items in memory during performance of the SOT, a high-demand WM task requiring substantial cognitive control, was related to the relatively pure measure of the number of items an individual can hold in visual short-term memory provided by WM capacity estimates from change detection tasks.

At the core of this question is a broader one about the cognitive processes that underlie various WM tasks. For example, abundant evidence points to a capacity limit in humans of approximately 4 items that can be concurrently held in WM, which is thought to be relatively directly tapped by canonical change detection tasks (Cowan, 2000; Lin & Luck, 2012; Luck & Vogel, 1997; Vogel & Machizawa, 2004). It is natural to ask whether this capacity limit constrains performance on WM tasks other than change detection tasks, especially those that are more complex and impose additional demands on cognitive control. One type of WM task that requires substantial cognitive control, known as complex span tasks, can also provide an estimate of WM capacity (or span), but psychometric studies indicate that capacity estimates from these tasks are distinct from estimates from change detection tasks; they load on separate factors and they exhibit different patterns of predictive validity, particularly with respect to measures of fluid intelligence (Conway, Cowan, Bunting, Theriault, & Minkoff, 2002; Cowan et al., 2005; Shipstead, Redick, Hicks, & Engle, 2012). Thus, the existing literature on complex span and change detection tasks suggests that there are at least two distinct sets of cognitive processes tapped by different classes of putative WM tasks, and it remains an open question as to whether the types of high-demand WM tasks used to tap cognitive control processes load on the same underlying cognitive capabilities required for successful performance of complex span tasks, as well as whether these tasks are constrained by the WM capacity limits tapped by change detection tasks.

Naturally, these two possibilities are not mutually exclusive, and so to examine the overall structure of relationships between WM tasks, we conducted a study in which a large sample of participants performed both complex span and change detection tasks, as well as an experimental WM task with a memory demand exceeding human WM capacity, and with a considerable demand on cognitive control; namely, the SOT. We elected to use the SOT as the third task in this study because our work on functional brain imaging with this task has demonstrated that within a single task trial, multiple (i.e. two or three) distinct brain systems are deployed at different points of a single trial in order to subserve task performance (Van Snellenberg et al., submitted). Furthermore, one of these brain systems includes regions that have been shown to be active during the performance of complex span tasks—namely bilateral dorsolateral prefrontal cortex, pre-supplementary motor areas (often called anterior cingulate cortex), and posterior parietal cortex (Chein, Moore, & Conway, 2011; Osaka et al., 2004; Smith et al., 2001)—as well as in the intraparietal sulcus and lateral occipital regions that have been shown to be active during change detection tasks and whose activation is correlated with capacity estimates from these tasks (Todd & Marois, 2004, 2005; Xu & Chun, 2006). Given the overlap in brain regions between the SOT and both complex span and change detection tasks, it seemed an ideal choice for probing whether an experimental WM task with high demands on cognitive control relies on the same core cognitive abilities as complex span and/or change detection tasks.

Our primary goals for this study were twofold, and largely exploratory. First, we hoped to develop estimates of WM capacity, which are notably different from simple accuracy measures, for the SOT and complex span tasks in a manner similar to the estimates obtained from change detection tasks (Cowan, 2000; Pashler, 1988; Rouder et al., 2008). Critically,

while canonical change detection tasks reflect a relatively pure measure of the number of items that can be simultaneously represented by an individual, abundant evidence indicates that performance on complex span tasks reflects a broader range of cognitive processes associated with encoding, maintenance, and retrieval, as well as selective attention and interference control (e.g., Cowan et al., 2005; Shipstead et al., 2012). The goal here was to evaluate whether a simple capacity measure for complex span tasks, as well as the SOT, would at least in part reflect the same core WM capacity estimated from change detection tasks; if so, capacity measures from these tasks should exhibit at least a modest positive correlation with those from a change detection task. Second, we sought to examine the pattern of correlations between, and the factor structure of, the three tasks discussed above in order to better understand how and whether multiple core cognitive capabilities (the putative latent variables subserving task performance) drive performance on different types of WM tasks—namely, change detection, complex span, and SOT tasks—and/or are related to an indicator of fluid intelligence, SAT scores.

Method

Participants

All procedures were approved by the Columbia University Morningside IRB. Informed consent to participate in the study was obtained from 121 undergraduates at Columbia University, who received either remuneration or course credit for their participation. This sample size was selected a priori to achieve 90% power to detect pairwise correlations of 0.3, which was based on a previously reported correlation between reading span and a change detection task (Cowan et al., 2005). The mean (SD) age of participants was 21.2 years (3.82), with a range of 18 - 35 years, and 67 participants (55.3%) were female. Data from three participants were excluded because of technical issues, falling asleep during testing, and “taking notes” during the reading span task, respectively. Reading span data from an additional 11 participants were excluded due to technical issues in one case, having English as a second language in two cases, and failing to achieve 85% correct on the secondary task in eight cases (a standard procedure for complex span tasks; see Conway et al., 2005). Change detection data from an additional four subjects were excluded due to technical issues. SAT scores were self-reported by 67 of the participants; the remainder of the participants either had not taken the SAT, could not remember their scores, or refused to provide them. In summary, SOT data were available for 118 participants, change detection data were available for 114 participants, and complex span data were available for 107 participants. This resulted in complete data on all three WM tasks for 101 participants, and data on all three WM tasks and SAT scores for 58 participants.

Behavioral Testing Procedures

Participants were tested on all behavioral tasks in a lit sound-attenuated booth on a PC desktop computer with a 17" LCD monitor. Participants were observed by a research assistant from outside of the booth through a glass door while they completed the tasks. Participants completed several practice trials of all tasks to ensure comprehension of task instructions and eliminate early practice effects. Tasks were presented in a counterbalanced order, and following completion of all three tasks, participants were asked to type their SAT score (and the year the score was obtained) into a Matlab dialog box, associated only with their sequentially-assigned subject number. SAT scores from tests taken prior to 2006 were multiplied by 1.5 in order to adjust for the change in the maximum score of the test in that year.

Change-detection task—Participants completed three blocks of 48 trials each of the change detection task, which was modeled after Luck and Vogel's (1997) classic study. On

each trial, participants were shown a display of either four or eight colored squares—a pseudo-random combination of red (RGB color 255, 0, 0), green (0, 255, 0), blue (0, 0, 255), yellow (255, 255, 0), white (255, 255, 255), and black (0, 0, 0)—on a grey background (half of the trials in each block contained four items while the other half contained eight, presented in pseudo-randomized order) for 500 ms, followed by a fixation cross for 1000 ms (the fixation was visible throughout each trial, but not during the inter-trial interval; ITI). Following this, a single square was presented in the same location as one of the stimuli from the target display, either in the same color as the original presentation or a different color (on 50% of the trials, in pseudo-randomized order). Participants then had up to 5 s to respond with a button press on a computer keyboard to indicate whether the probe matched or did not match the color of the corresponding item in the target display. After a response was made there was a 1 s ITI before the start of the next trial.

Stimuli subtended approximately 1.43° of visual angle, and could appear anywhere within an area in the middle of the screen covering 23.38° of visual angle in width and 17.47° of visual angle in height (two-thirds of the total screen size), with the exception of a 1.43° band at the center of both the X and Y axes defined by the fixation cross (that is, no stimulus could appear at the horizontal or vertical center of the display). This area was divided into quadrants and for 4-item arrays one stimulus was placed at a random location in each quadrant, while for 8-item arrays two stimuli were placed at a random location in each quadrant with at least 2.15° between the center of each stimulus in both the horizontal and vertical direction.

Self-ordered working memory task—Participants completed 24 trials of the SOT, with each trial containing eight steps on which a response was required. At the start of each trial, eight simple line drawings of three-dimensional objects were presented in a three-by-three grid, with the central position of the grid empty. Stimuli were the same as those used by Curtis, Zald, and Pardo (2000); unique stimuli were used on each of the first 12 trials, and stimuli were repeated exactly once during the latter 12 trials. On each step, subjects had 7 s to move a mouse cursor to select any object that had not been selected on a previous trial (thus, all responses are correct on the first step). Once a selection was made, a white outline was displayed around the selected object until nine seconds had elapsed from the start of the step. At this point the objects in the display were pseudo-randomly rearranged in the grid, with the blank space appearing in the same location as the most recently selected item (to prevent participants from using a spatial strategy or simply responding in the same location on each trial). If no response was made in seven seconds, a white outline was displayed for two seconds around a randomly selected object that would have been a correct response; participants were instructed to remember this object as if they had selected it themselves. If an incorrect selection was made, a red box was displayed over the object until seven seconds had elapsed from the start of the step, after which the same procedure was followed as in the case when a participant made no response. The ITI was 9 seconds. Stimuli were up to 7.15° of visual angle horizontally and vertically, but most objects were slightly smaller than that in one direction. The grid was arranged such that there was 0.36° of visual angle between the outermost edges of each of the stimuli.

Complex span task—The complex span task used was the automated reading span task available at <http://psychologygatech.edu/renglelab/Eprime1.html> (see Unsworth, Heitz, Schrock, & Engle, 2005). Participants completed 15 trials of the task, including three trials each of all set sizes between three and seven items. For each item in a trial, participants were presented with a grammatical sentence that either made sense (e.g., “John left to go to the store”) or did not make sense (e.g., “John left to go to the toothbrush”). Participants clicked once with the mouse to indicate that they had read the sentence and then made a judgment as to whether or not the sentence made sense, and finally were presented with a letter that they

were required to remember. After all items within a trial had been presented, participants were given unlimited time to select all of the presented items in the correct order from a grid of 12 letters. Participants could also select a 'blank' item if they did not remember the item in a particular position. The time participants were given to read the sentences was determined by the mean time each individual participant took to read sentences during the practice session, plus 2.5 SD (see Unsworth et al., 2005 for justification of the equivalent procedure for the operation span task).

Estimation of Working Memory Capacity

Change-detection task—A large body of work has employed a simple model of WM to estimate capacity from change detection tasks, based on the assumption that participants have a fixed WM capacity that is fully utilized on every trial. That is, on a trial in which a

change occurs, the probability of the subject detecting the change is presumed to be $\frac{k}{N}$, where k is their WM capacity and N is the number of items in the display and $k \leq N$. Thus,

the change is *not* detected by the subject with probability $\frac{N-k}{N}$, but in these cases, the subject may still correctly guess that a change occurred with probability g , even though they were unable to make a correct determination from memory. Thus, the probability of the

subject making a hit (in signal detection terms) is $H = \frac{k}{N} + g \frac{N-k}{N}$. Conversely, the probability of a correct rejection on trials in which no change occurs is

$CR = \frac{k}{N} + (1-g) \frac{N-k}{N}$. By combining these two formulae and solving for k , one obtains $k = N(H + CR - 1)$, and so k can be estimated from the number of hits and correct rejections (note that the unknown parameter g drops out of the formula for k ; see Cowan, 2000, for the original specification of this formula, and see Pashler, 1988, for an equivalent formula for tasks using a whole-display probe rather than a single-item probe as used in the present study).

There are, however, two important problems with this simple form of the model. First, the model predicts perfect performance whenever $N = k$, which becomes especially problematic with smaller set sizes as it does not account for cases in which participants make an error due to reasons unrelated to WM capacity; for example, errors in executing the correct motor response or lapses in attention. Second, the model provides different estimates of WM capacity for a given participant at each investigated set size, which is inappropriate for a quantity that is presumed to be fixed. Rouder and colleagues (2008) dealt with the first issue by expanding the model to include an attention parameter. That is, they assume that on some trials participants fail to attend to the target stimuli and, as a result, no items are encoded

into WM. Thus, given the attention parameter a and the probability $d = \min\left(\frac{k}{N}, 1\right)$ that the probed item is in working memory, the probabilities of a hit or correct rejection in this model are:

$$H = a(d + g(1-d)) + (1-a)g, \text{ and} \\ CR = a(d + (1-g)(1-d)) + (1-a)(1-g).$$

This expanded model can no longer be solved directly for k , as there are two additional unknown parameters a and g , which do not drop out of the formula as g does in the simpler model. Rouder and colleagues dealt with this issue in a way that also addresses the second criticism of the simpler model outlined above, by using MLE to fit all three unknown

parameters k , a , and g simultaneously, thereby obtaining only a single estimate of k irrespective of the number of set sizes used in the study.¹

Self-ordered working memory task—A similar approach to that described for the change detection task can be applied to the SOT in order to obtain capacity estimates, however a substantively different model must be considered because of the different structure of the SOT. Under the same assumption of a fixed WM capacity that is fully utilized provided the participant is attending to the task, a participant making a response on step S of an SOT with a display of N items will have $m = \min(S - 1, k)$ items maintained in WM (and consequently $N - m$ items not maintained in WM). If $C_i = 1$ when the i th item has been previously selected and $C_i = 0$ when it has not, the probability that the participant will select the i th item is:

$$P(\text{select}_i) = C_i \frac{1 - \frac{m}{S-1}}{N - m} + (1 - C_i) \frac{1}{N - m}.$$

Because an error can only occur if the participant selects an item that was previously selected, we need only be concerned with the first ratio, and the C_i term drops out as it is always equal to one in this case. Thus, the probability of an error is simply the sum of the probabilities of selecting each of the previously selected stimuli, and because these probabilities are equal to each other and there are $S - 1$ previously selected stimuli at any step of the task, the probability of an error becomes:

$$E = \sum_{i=2}^S \frac{1 - \frac{m}{S-1}}{N - m} = (S - 1) \frac{1 - \frac{m}{S-1}}{N - m} = \frac{S - 1 - m}{N - m}.$$

It is critical to note that when $S - 1 = k$, $m = S - 1$, and so the probability of an error becomes 0. Thus, similar to the change detection task, if a participant makes an error at an early step because of a lapse of attention or a motor error, the model is forced to presume that their capacity is very low. For this reason, it is necessary to include an attention parameter in the model, as with the change detection task. However, while it is straightforward to specify how participants behave during an attention lapse in the two-alternative-forced-choice structure of the change detection task, it is less clear what occurs in the SOT. That is, it seems unlikely that in the event of a lapse of attention during a trial, participants would guess randomly amongst all available stimuli and then resume normal responding on subsequent steps. Consequently, we treated the lapse parameter as a simple probability of an error for reasons unrelated to a participant's WM capacity. Consequently, the probability of an error becomes simply:

$$E = a \frac{S - 1 - m}{N - m} + (1 - a).$$

With this model, MLE can be used as with the change detection task to estimate individuals' WM capacity, with k and a as free parameters to be fit by the model.

Complex span task—The model needed to obtain MLE estimates of WM capacity from complex span tasks is somewhat more complex than for change detection and SOT tasks,

¹Although they do not spell out the MLE procedure used, the approach taken here was to treat each trial as a Bernoulli trial with probability determined by the above formulae and use a brute-force search of possible values for all three parameters.

given the complex nature of the task and how behavioral responses are made on each trial. First, assuming the participant is fully attending to the trial, the probability that any given item on a trial is maintained in WM in a fixed-capacity model is the same that for change

detection; that is, $d = \min\left(\frac{k}{N}, 1\right)$. However, in the case where an item is not maintained in WM the participant may either guess with (unknown) probability b from one of the twelve items in the display, or they may elect to make no response (i.e. by responding with a 'blank') with probability $1 - b$. Assuming they attempt to guess, they have probability

$g_a = \frac{1}{12 - c}$ of guessing correctly, where c is the number of items they can eliminate from the full range of possibilities based either on items they do remember or based on the number of items they have already selected by guessing. At a minimum, $c_{min} = \max[R, \min(k, N)]$, where R is the number of letters that have already been selected in the course of responding to the trial. At a maximum, c is somewhat more complex as the participant may be holding the final k items in memory and guessing on the first $N - k$ items, in which case $c_{max} = \min(k, N) + \min(R, N - k - 1)$. Note that when $R = 0$, $c_{min} = c_{max} = \min(k, N)$. While c itself is unknown on any given trial, under the assumption that the items a participant is able to hold in WM are randomly distributed amongst all of the N items, the expected value of c is

simply the average of c_{min} and c_{max} , that is $E(c) = \frac{c_{min} + c_{max}}{2}$. Consequently, we use this expected value to estimate c for each response of each trial. In this model, then, the probability of a participant being correct on any particular item on a given trial is:

$$I_{correct} = d + b(1 - d)g_a.$$

As with the change detection and SOT tasks, however, this model cannot accommodate errors due to reasons unrelated to WM capacity, such as inattention. Consequently, we must introduce the parameter a as with the previous tasks, such that participants attend to an item with (unknown) probability a and fail to attend with probability $1 - a$. In this latter case, they may again guess amongst the remaining items with probability b (for simplicity, we assume this is the same probability with which they will attempt to guess when they attend to an item but are unable to maintain it in WM due to capacity constraints), or they may simply elect to make no response with probability $1 - b$. Assuming they guess, they have

probability $g = \frac{1}{12 - R}$ of guessing correctly, under the assumption that the contents of WM are momentarily not available to them, and so they can only eliminate previously selected items from the full set of items from which they can make a response. In this larger model, then, the probability of a participant being correct on any particular item on a given trial is:

$$I_{correct} = a(d + b(1 - d)g_a) + g(1 - a)b.$$

In addition, the probability of the participant making no response on any given item is:

$$I_{NR} = a(1 - d)(1 - b) + (1 - a)(1 - b).$$

With these formulae, the probability of a correct response, incorrect response, or no response can be used to fit the model parameters k , a , and b to the model using MLE.

Assessment of Capacity Model Fits and Reliabilities

In order to evaluate the fit of the capacity models for each of the three WM tasks, we followed the approach of Rouder et al. (2008) in formally testing the null hypothesis that a perfectly specified model (what might be called an omniscient model—one in which performance at each set size is fit with its own free parameter so that the predicted probability of a correct response is the observed proportion of correct responses) provides no additional information over the capacity model, using a log-likelihood goodness-of-fit test on each individual subjects' data.

Furthermore, to test whether the capacity models provide reliable estimates of capacity, we calculated capacity estimates on two halves of each participants data (split-half reliability), and adjusted the resulting correlation with the Spearman-Brown prediction formula to obtain an estimate of the reliability of capacity models from each task. Because reliability may be heavily dependent on precisely which trials end up in each half, we randomly selected half of the trials in each task 100 times and report the median reliability for each of these 100 split-half reliability estimates (more than 100 analyses were not conducted because the model fitting procedures are very computationally intensive, and need to be carried out twice per subject for each reliability estimate).

Correlation Analyses

Initial examination of the data revealed that scores from the SOT task were decidedly non-normal (SOT: skew = -2.41, kurtosis = 7.53; complex span: skew = -0.33, kurtosis = -0.33; change detection: skew = 0.18, kurtosis = -0.47). As a result, we opted to use a non-parametric bootstrapping method to obtain confidence intervals (CI) and *P* values for all statistical tests of interest. The bootstrapping method employed was the *bias-corrected and accelerated bootstrap* (BC_a; Efron & Tibshirani, 1993)², using 10,000 bootstrap iterations for each analysis.

Given that this is the first study to attempt to develop capacity estimates for either the SOT or complex span tasks, one potential concern with using only capacity estimates to evaluate whether these tasks share at least some underlying cognitive processes is that a lack of correlation between capacity estimates from the SOT and complex span tasks may simply indicate that a capacity model is inappropriate for these tasks. Thus, in order to ensure that the pattern of correlations observed between capacity estimates for each of the tasks is not simply due to an inappropriate capacity model for one or more of the tasks, we also computed pairwise correlations of accuracy (percent correct) for each WM load in each of the three tasks—a total of 59 pairwise correlations.

Exploratory Factor Analysis

In order to evaluate whether underlying latent variables may drive performance on some of the WM tasks and SAT scores, we also carried out an exploratory factor analysis (EFA) with promax rotation. In order to determine the number of factors to retain in the model, we first carried out a principal components analysis and used Horn's parallel analysis (PA; e.g., Hayton, Allen, & Scarpello, 2004) to determine the number of factors to retain. However, when this was carried out on the capacity estimates from each task, no factors at all were retained by this method. We presumed that this was due to the very small number of

²While Efron and Tibshirani do not provide a means of calculating BC_a *P* values directly, this can be done by determining α for the $(1 - \alpha)\%$ CI that would have its upper or lower bound at exactly the null-hypothesis value being tested. Thus, one simply observes the proportion of the bootstrap distribution falling below the null-hypothesis value of the statistic and applies the function inverse (i.e., given the function $f(x)$, the function inverse of $f(x)$ is $f^{-1}(x)$ such that $f^{-1}(f(x)) = x$) of the BC_a correction given for CIs to this proportion.

observed variables in this analysis (four in total), and so opted to carry out the analysis on the raw accuracy data from each set size of each task, as well as SAT. This resulted in 15 observed variables (7 for steps two through eight of the SOT, two for the change detection task, five for the complex span task, and SAT). Furthermore, because several of the accuracy values were again non-normal (considerable ceiling effects at lower set sizes in all tasks), we used the BC_a bootstrap to determine P values for the factor loadings of each variable. Again, 10,000 bootstrap iterations were carried out, and to ensure that factor ordering remained constant on each iteration, we compared the resulting factor loadings to the observed factor loadings and reordered the factors so that the bootstrapped loadings were always associated with the observed factor to which they were the most similar.

Results

Capacity Model Fits and Reliability

The model fits for estimates of WM capacity for each of the three tasks were generally good, with the goodness-of-fit test failing to reject the null hypothesis of no improvement in model fit by an omniscient model in 76% of participants in the SOT, 77% of participants in the change detection task, and 89% of participants in the complex span task. Reliability estimates, however, were poor for both the change detection and complex span tasks, at 0.32 and 0.45 respectively, while the SOT fared much better at 0.77. Because the maximum population correlation between two variables is the square root of the product of their reliabilities, this implies that the maximum correlations that could be observed between these tasks are quite modest (in the range from 0.38 to 0.59, depending on the task).

Another important metric of the appropriateness of model fits is the value of the attention parameter (a) included in each of the capacity models; values of a that are very low would indicate either that participants were not attending to the task appropriately or that the model was treating most of the errors by participants as being due to inattention rather than low WM capacity. The mean (SD) of values of a for each of the three tasks were 0.955 (0.035) for the SOT, 0.857 (0.127) for the change detection task, and 0.943 (0.080) for the complex span task. This value of a for the change detection task is unusually low, as higher values are typically observed (e.g., Gibson, Wasserman, & Luck, 2011). This is probably attributable to the fact that we did not include trials with a very small number of items (i.e. 1 or 2), which have been used in other studies and help to constrain the value of a by providing a set of trials in which essentially all errors can be attributed to attention lapses. In addition, the attention parameters were correlated for model fits of the SOT and change detection task ($r = 0.29$; $P = 0.012$) and showed a trend level correlation for the change detection and complex span tasks ($r = 0.18$; $P = 0.074$), but were not correlated between the SOT and complex span task ($r = 0.10$; $P = 0.298$). These results indicate that the capacity models were consistent with participants attending to a large majority of trials, at least for the SOT and complex span task. The values of the a parameter for the change detection task were somewhat lower however, with participants on average not attending to over 14% of trials.

Task Performance

Performance data (percent correct) at each load is displayed in Figure 1 for all three tasks. The mean (SD) value of k was 6.20 (1.08) for the SOT task, 4.39 (1.09) for the change detection task, and 5.36 (1.14) for the complex span task; capacity estimates for all three tasks were all significantly different from each other (all $P < 0.001$).

Correlation Analyses

The correlation between capacity estimates in the SOT and change detection tasks was 0.28 ($P < 0.001$) with a 95% CI of 0.11 - 0.47. The correlation between capacity estimates in the

SOT and complex span task was -0.15 ($P = 0.086$) with a 95% CI of $-0.31 - 0.02$. Finally, the correlation between capacity estimates in the change detection and complex span tasks was -0.05 ($P = 0.600$) with a 95% CI of $-0.23 - 0.14$.

The correlation between SOT and SAT was 0.08 ($P = 0.464$) with a 95% CI of $-0.16 - 0.30$, between change detection and SAT was 0.01 ($P = 0.954$) with a 95% CI of $-0.28 - 0.25$, and between complex span and SAT was 0.12 ($P = 0.345$) with a 95% CI of $-0.15 - 0.35$.

Amongst the 35 correlations between the seven loads of the SOT and five loads of the complex span task, none achieved statistical significance after false discovery rate (FDR) correction for multiple comparisons (Benjamini & Hochberg, 1995), and only two achieved significance at an uncorrected alpha of $P < 0.05$. These were the correlation between performance at step 6 of the SOT and a 3-item load in the complex span task ($r = 0.25$), and the correlation between performance on step 8 of the SOT and a 4-item load in the complex span task ($r = 0.18$). Given that with 35 correlations the expected number of Type I errors is 1.75, this strongly suggests no relationship between performance on these two tasks. The 10 correlations between loads of the change detection and complex span tasks were similar, with no correlations passing FDR correction and only one correlation significant without correction: the correlation between performance on 8-item sets in the change detection task and 3-item sets on the complex span task ($r = 0.20$). In contrast, of the 14 correlations between performance at various loads of the SOT and the change detection task, 10 of them passed FDR correction (r ranged from $0.24 - 0.50$). The four correlations not passing FDR correction (or an uncorrected alpha of 0.05) were the correlations between steps 2 through 5 of the SOT and the 8-item load of the change detection task. Thus, the pattern of pairwise correlations between accuracy on each of the tasks broadly supports the results of the analyses using capacity estimates as the sole outcome measure for each task, demonstrating that the pattern of results obtained from the capacity estimates is not an artifact of poor or inappropriate modeling of WM capacity for each of the three tasks.

Exploratory Factor Analysis

The PA factor retention decision resulted in two factors; factor loadings are shown in a biplot in Figure 2. All of the SOT accuracy measures and accuracy on 4-item displays of the change detection task significantly loaded on Factor 1 (all $P < 0.05$), while complex span accuracy measures for five to seven items significantly loaded on Factor 2 (all $P < 0.05$) and complex span accuracy measures for three and four items showed a trend towards significant loading on Factor 2 ($P = 0.057$ and $P = 0.092$, respectively). Factor loadings, confidence intervals, and P values are presented for all variables in Table 1. It is also noteworthy that 8-item change detection task accuracy and SAT scores approached marginal significance ($P < 0.2$) for Factors 1 and 2, respectively; as noted in the methods, inclusion of SAT drastically reduced our available sample size because only approximately half of our sample was able to report SAT scores. A follow-up EFA excluding SAT in order to retain more of our sample (101 participants total) produced significant loadings on Factor 1 for all of the SOT and change detection measures, and a significant loading on Factor 2 for all of the complex span measures (all $P < 0.01$).

Discussion

The results of both the pairwise correlations between tasks and the EFA strongly suggest that there are at least two distinct cognitive abilities underlying capacity estimates and performance on the three WM tasks studied here; based on previous work, it is likely that capacity on complex span tasks as well as performance on the SAT is driven largely by cognitive abilities closely linked to fluid intelligence (see Cowan et al., 2005)—likely including active maintenance of information in WM in the face of distraction from a

secondary task and controlled retrieval of information when active maintenance fails (see Unsworth & Engle, 2007)—while capacity on the SOT and change detection tasks is driven by a limit on the number of items individuals can hold in the focus of attention at the same time (Cowan, 2000). Thus, WM and WM capacity are not necessarily unified concepts as used in the literature, and depending on the task under discussion may refer to one of at least two distinct cognitive abilities.

In addition, this study presents methods for obtaining capacity estimates from two tasks for which such procedures were not previously available, the SOT and complex span, and provides a reliability estimate for capacity estimates from the change detection task. In general, the models of WM capacity presented here do a good job of explaining behavioral performance for most participants. However, reliability for the change detection task was surprisingly poor, which is likely due to the fact that we did not include any trials with a very low set size, which would have given a more stable estimate of the attention parameter in the capacity model, and consequently more stable estimates of capacity. Indeed, when these data were analyzed without an attention parameter, using the capacity formula given in Cowan (2000), substantially better reliability was achieved (Van Snellenberg, 2012) and would likely have been even higher with the inclusion of low set size trials. The results of these analyses without an attention parameter were also broadly consistent with those reported here (*ibid.*), indicating that inclusion of this parameter did not unduly impact the results of the present study. Another consideration is that change-localization tasks (e.g., Gold et al., 2006), which are essentially change detection tasks in which a change occurs on every trial and participants must instead indicate at which location in the array a change occurred, have been shown to have greater reliability than change detection tasks (Johnson et al., 2013; Kyllingsbaek & Bundesen, 2009), and so may be preferable to change detection tasks in future work. Reliability for the complex span task was also quite poor, possibly due to the inclusion of only 15 trials in the task—one implication of this finding is that researchers wishing to use capacity estimates for complex span tasks would benefit from employing more trials, for example the Spearman-Brown prediction formula estimates that reliability with 45 trials would be 0.71, much stronger than the 0.45 we observed with 15 trials.

It is also worth noting that the capacity estimates for the SOT, while reasonably well fit and fairly reliable, were much higher than would be expected for a true estimate of the number of items individuals can hold in the focus of attention (i.e. 6.2 items compared to 4.4 for the change detection task). While this may to some extent be due to sample characteristics (predominantly Columbia University undergraduates), it is almost certain that additional cognitive strategies can be brought to bear to subserve task performance; indeed, other work with this task strongly suggests that this is the case (Van Snellenberg et al., submitted). Nonetheless, the modest correlation we observed between SOT and change detection capacity indicates that performance on the SOT is at least constrained to some degree by ‘pure’ WM capacity. However, it is worth noting that whatever strategies individuals use to improve their estimated capacity beyond the usual ‘four plus or minus one’ limit, the present data suggest that they do not depend on the same cognitive abilities tapped by complex span tasks.

Moreover, it seems that the number of items that individuals can maintain in memory in the complex span task employed here is entirely unrelated to the number of items individuals can maintain in the change detection or SOT tasks. This implies that performance on complex span tasks is not constrained simply by the number of items that can be held in the so-called focus of attention, but may have more to do with the amount of information that can be maintained in the face of the distraction inherent in performing a secondary, unrelated task (Unsworth & Engle, 2007); a feature of complex span tasks that is believed to

be critical to its validity in predicting other measures of fluid intelligence (Conway et al., 2005). What the present data make entirely clear is that this cognitive ability is unrelated to the capacity-limitations that underlie the change detection task and the broadly construed cognitive control required by at least the SOT, if not other complex experimental WM tasks.

Limitations

Several caveats are important to consider when interpreting the results of the present study. First, our sample was primarily made up of undergraduates from an elite university, thereby limiting the generalizability of our results to the general population. Indeed, estimates of capacity in each of our tasks were quite high (all >4). Consequently, it is entirely possible that the strength of correlations between tasks was limited by restriction of range in the general cognitive abilities of our participants, or even that a relatively high performing sample such as ours may bring different cognitive capabilities to bear on performing these tasks than would a lower-performing sample. Second, we used only a single task for each type of WM task under study, meaning that these results may be specific to the tasks used rather than being a general property of complex span tasks, change detection tasks, and complex WM tasks with heavy demands on cognitive control. Third, conclusions drawn from our correlational results are necessarily subject to all of the usual caveats related to correlations, such as the existence of third variables. While it is unlikely that a variable such as fatigue or motivation produced the correlation between change detection and SOT, given that these tasks were not also correlated with complex span (which one would expect to be subject to effects of these variables as well), it does remain possible that some other cognitive capability besides WM capacity induced the observed correlation between these tasks. Finally, the conclusions that can be drawn from any of the analyses of SAT scores must be strongly tempered, given that these scores were self-reported (the Columbia University admissions office refused to release scores to us, even with permission from our participants) and were available in only approximately half of our sample.

Another major issue stems from the capacity model used in this study for the complex span task. Because complex span requires memory for serial position of items in addition to the items themselves, it is likely that errors occur in this task for reasons other than simple WM capacity constraints. While for the purposes of the present study we attempted to develop a measure of capacity that was as analogous as possible to the models used for change detection and the SOT, it would require substantial further work to validate this model as an appropriate outcome measure for complex span tasks in other studies; for example, by showing that the capacity estimates from this model correlate as well as standard measures with estimates of fluid intelligence. While the pattern of results we observed with this model was corroborated by analyses using simple accuracy measures, it remains entirely possible that it is not a generally appropriate estimate of WM capacity.

Acknowledgments

This work was supported by NIMH 5P50 MH086404. The authors would like to thank Melanie Pincus for her assistance in setting up the study protocol, and Debbie Fraser, Mona Griffin, and Serena di Stefani for their assistance in running participants.

References

- Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995; 57:289–300.
- Chein JM, Moore AB, Conway AR. Domain-general mechanisms of complex working memory span. *Neuroimage*. 2011; 54:550–559. [PubMed: 20691275]

- Conway ARA, Cowan N, Bunting MF, Theriault D, Minkoff S. A latent variable analysis of working memory capacity, short term memory capacity, processing speed, and general fluid intelligence. *Intelligence*. 2002; 30:163–183.
- Conway ARA, Kane MJ, Bunting MF, Hambrick DZ, Wilhelm O, Engle RW. Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*. 2005; 12:769–786.
- Cowan N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*. 2000; 24:87–185. [PubMed: 11515286]
- Cowan N, Elliott EM, Saults JS, Morey CC, Mattox S, Hismjatullina A, Conway ARA. On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*. 2005; 51:42–100. [PubMed: 16039935]
- Curtis CE, Zald DH, Pardo JV. Organization of working memory within the human prefrontal cortex: a PET study of self-ordered object working memory. *Neuropsychologia*. 2000; 38(11):1503–1510. [PubMed: 10906375]
- Efron, B.; Tibshirani, RJ. *An Introduction to the Bootstrap*. Chapman & Hall/CRC; New York, NY: 1993.
- Gibson B, Wasserman E, Luck SJ. Qualitative similarities in the visual short-term memory of pigeons and people. *Psychonomic Bulletin & Review*. 2011; 18:979–984. [PubMed: 21748417]
- Gold JM, Fuller RL, Robinson BM, McMahon RP, Braun EL, Luck SJ. Intact attentional control of working memory encoding in schizophrenia. *Journal of Abnormal Psychology*. 2006; 115:658–673. [PubMed: 17100524]
- Hayton JC, Allen DG, Scarpello V. Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*. 2004; 7:191–205.
- Johnson MK, McMahon RP, Robinson BM, Harvey AN, Britta H, Leonard CJ, Gold JM. The relationship between working memory capacity and broad measures of cognitive ability in healthy adults and people with schizophrenia. *Neuropsychology*. 2013; 27:220–229. [PubMed: 23527650]
- Kyllingsbaek S, Bundesen C. Changing change detection: Improving the reliability of measures of visual short-term memory capacity. *Psychonomic Bulletin & Review*. 2009; 16:1000–1010. [PubMed: 19966249]
- Lin P-H, Luck SJ. Proactive interference does not meaningfully distort visual working memory capacity estimates in the canonical change detection task. *Frontiers in Psychology*. 2012; 3:42. [PubMed: 22403556]
- Luck SJ, Vogel EK. The capacity of visual working memory for features and conjunctions. *Nature*. 1997; 390(6657):279–281. doi: 10.1038/36846. [PubMed: 9384378]
- Osaka N, Osaka M, Kondo H, Morishita M, Fukuyama H, Shibasaki H. The neural basis of executive function in working memory: An fMRI study based on individual differences. *Neuroimage*. 2004; 21:623–631. [PubMed: 14980565]
- Pashler H. Familiarity and visual change detection. *Perception & Psychophysics*. 1988; 44:369–378. [PubMed: 3226885]
- Rouder JN, Morey RD, Cowan N, Zwilling CE, Morey CC, Pratte MS. An assessment of fixed-capacity models of visual working memory. *Proc Natl Acad Sci U S A*. 2008; 105:5975–5979. [PubMed: 18420818]
- Shipstead Z, Redick TS, Hicks KL, Engle RW. The scope and control of attention as separate aspects of working memory. *Memory*. 2012; 20:608–628. [PubMed: 22734653]
- Smith EE, Geva A, Jonides J, Miller P, Reuter-Lorenz P, Koeppel RA. The neural basis of task-switching in working memory: Effects of performance and aging. *Proc Natl Acad Sci U S A*. 2001; 98:2095–2100. [PubMed: 11172081]
- Smith EE, Van Snellenberg JX. Capacity and processing deficits in working memory in schizophrenia. *Schizophrenia Bulletin*. 2011; 37:228.
- Todd JJ, Marois R. Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*. 2004; 428(6984):751–754. doi: 10.1038/nature02466. [PubMed: 15085133]
- Todd JJ, Marois R. Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cognitive, Affective, and Behavioral Neuroscience*. 2005; 5:144–155.

- Unsworth N, Engle RW. The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*. 2007; 114:104–132. [PubMed: 17227183]
- Unsworth N, Heitz RP, Schrock JC, Engle RW. An automated version of the operation span task. *Behavior Research Methods*. 2005; 37:498–505. [PubMed: 16405146]
- Van Snellenberg, JX. Dissertation. Columbia University; 2012. An Investigation of the Neural Correlates of Working Memory in Healthy Individuals and Individuals With Schizophrenia. Permanent URL, <http://hdl.handle.net/10022/AC:P:13157>
- Van Snellenberg JX, Girgis RR, Read C, Thompson JL, Weber J, Wager TD, Smith EE. Individuals with schizophrenia fail to show normative inverted-U activation in response to fine-grained working memory load manipulation. *Schizophrenia Bulletin*. 2013; 39:S251–252.
- Van Snellenberg, JX.; Slifstein, M.; Read, C.; Weber, J.; Thompson, JL.; Wager, TD.; Smith, EE. Dynamic shifts in brain network activation during working memory task performance. (submitted)
- Van Snellenberg JX, Wager TD, Abi-Dargham A, Urban N, Smith EE. Parametric variation in working memory demand in patients with schizophrenia: A behavioral and neuroimaging pilot study. *Biological Psychiatry*. 2010; 67:158S.
- Vogel EK, Machizawa MG. Neural activity predicts individual differences in visual working memory capacity. *Nature*. 2004; 428(6984):748–751. doi: 10.1038/nature02447. [PubMed: 15085132]
- Xu Y, Chun MM. Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*. 2006; 440:91–95. [PubMed: 16382240]

Figure 1. Behavioral performance on the A) self ordering working memory task, B) change detection task, and C) complex span task. Error bars reflect ± 1 standard error.

Figure 2.

Biplot of factor loadings from the exploratory factor analysis. SOT = self ordering working memory task; CD = change detection task; CS = complex span task.

Table 1

Exploratory Factor Analysis Loadings

Variable	Factor 1			Factor 2		
	Factor Loading	95% CI	P	Factor Loading	95% CI	P
SOT Step 2	0.39	0.03 – 0.65	0.040	–0.05	–0.65 – 0.35	0.649
SOT Step 3	0.58	0.17 – 0.88	0.007	–0.16	–0.64 – 0.23	0.262
SOT Step 4	0.73	0.42 – 0.91	<0.001	–0.10	–0.41 – 0.11	0.217
SOT Step 5	0.78	0.41 – 0.94	<0.001	–0.11	–0.46 – 0.21	0.222
SOT Step 6	0.88	0.72 – 0.96	<0.001	0.02	–0.18 – 0.26	0.994
SOT Step 7	0.91	0.71 – 1.02	<0.001	0.03	–0.25 – 0.29	0.955
SOT Step 8	0.85	0.56 – 0.95	<0.001	0.03	–0.20 – 0.47	0.749
CD 4-Item	0.64	0.06 – 0.87	0.031	0.10	–0.33 – 0.84	0.769
CD 8-Item	0.33	–0.11 – 0.67	0.136	0.11	–0.51 – 0.75	0.762
CS 3-Item	0.12	–0.14 – 0.56	0.399	0.49	–0.02 – 0.99	0.057
CS 4-Item	0.05	–0.13 – 0.36	0.641	0.34	–0.07 – 0.77	0.092
CS 5-Item	0.08	–0.16 – 0.40	0.458	0.54	0.04 – 1.02	0.038
CS 6-Item	–0.09	–0.31 – 0.23	0.416	0.76	0.24 – 1.13	0.011
CS 7-Item	–0.02	–0.26 – 0.19	0.658	0.62	0.10 – 1.04	0.034
SAT	0.08	–0.26 – 0.39	0.528	0.25	–0.17 – 0.57	0.181

Note. CI = confidence interval. SOT = self-ordered working memory task. CD = change detection. CS = complex span.