Cogn Affect Behav Neurosci
DOI 10.3758/s13415-013-0235-x

Capacity estimates in working memory: Reliability and interrelationships among tasks

Jared X. Van Snellenberg · Andrew R. A. Conway · Julie Spicer · Christina Read · Edward E. Smith

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 as capacity estimates from at least one type of WM task—complex span—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, and may or may not refer to the same cognitive capability. We here develop maximum-likelihood models of capacity from each of these tasks—as well as from a third WM task that places heavy demands on cognitive control, the self-ordered WM task (SOT)—and show that the capacity estimates from change detection and complex span tasks are not correlated with each other, although capacity estimates 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 it applies to each of these three tasks.

Keywords Working memory · Short-term memory · Cognitive control

J. X. Van Snellenberg (✉)
Department of Psychiatry, Columbia University College of Physicians and Surgeons, New York, NY, USA
e-mail: jaredvs@gmail.com

J. X. Van Snellenberg · C. Read
Division of Translational Imaging, New York State Psychiatric Institute, New York, NY, USA

A. R. A. Conway
Department of Psychology, Princeton University, Princeton, NJ, USA

J. Spicer · E. E. Smith
Department of Psychology, Columbia University, New York, NY, USA

E. E. Smith
Division of Cognitive Neuroscience, New York State Psychiatric Institute, New York, NY, USA

Published online: 08 January 2014

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 the 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, 2001; 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, Girgis, et al., 2013; Van Snellenberg, Wager, Abi-Dargham, Urban, & Smith, 2010), he was eager to develop a similar capacity model for the task that 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 that an individual can hold in visual short-term memory that is 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 four items that can be concurrently held in WM, which is thought to be tapped relatively directly by canonical change detection tasks (Cowan, 2001; 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 have indicated that capacity estimates from these tasks are distinct from the 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, Therriault, & Minkoff, 2002; Cowan et al., 2005; Shipstead, Redick, Hicks, & Engle, 2012). Thus, the existing literature on complex span and change detection tasks suggests that at least two distinct sets of cognitive processes are tapped by different classes of putative WM tasks, and it remains an open question 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, 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 in the trial to subserve task performance (Van Snellenberg, Slifstein, et al., 2013). 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, presupplementary 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 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, the SOT 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, 2001; Pashler, 1988; Rouder et al., 2008). Critically, whereas 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 reflecting 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 Institutional Review Board. 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 .3, a value 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 of the participants (55.3 %) were female. The data from three participants were excluded because of technical issues, falling asleep during testing, and "taking notes" during the reading span task, respectively. The 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 participants were excluded due to technical issues. SAT scores were self-reported by 67 of the participants; the remainder of the participants 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 for 114 participants, and complex span data for 107 participants. This resulted in complete data on all three WM tasks for 101 participants, and data on all three WM tasks plus SAT scores for 58 participants.

Behavioral testing procedures

Participants were tested on all of the behavioral tasks in a lit, sound-attenuated booth on a PC desktop computer with a 17-in. LCD monitor. Participants were observed by a research assistant from outside of the booth through a glass door while they completed the tasks. The participants completed several practice trials of all tasks, to ensure comprehension of the task instructions and eliminate early practice effects. The 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 that the score was obtained) into a MATLAB dialog box associated only with their sequentially assigned participant 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 that in Luck and Vogel's (1997) classic study. On each trial, participants were shown a display of either four or eight colored squares—a pseudorandom 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 gray background (half of the trials in each block contained four items, whereas the other half contained eight, presented in pseudorandomized order) for 500 ms, followed by a fixation cross for 1,000 ms (the fixation was visible throughout each trial, but not during the intertrial 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 in the original presentation or a different color (on 50 % of the trials, again in pseudorandomized order). Participants then had up to 5 s to respond with a buttonpress 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, a 1-s ITI was presented before the start of the next trial.

The stimuli subtended approximately 1.43º of visual angle and could appear anywhere within an area in the middle of the screen covering 23.38º of visual angle in width and 17.47º of visual angle in height (two thirds of the total screen size), with the exception of a 1.43º band at the center of both the x- and y-axes that was defined by the fixation cross (i.e., no stimulus could appear at the horizontal or vertical center of the display). This area was divided into quadrants; for four-item arrays, one stimulus was placed at a random location in each quadrant, whereas for the eight-item arrays, two stimuli were placed at a random location in each quadrant, with at least 2.15º between the centers of the stimuli in both the horizontal and vertical directions.

Self-ordered WM task The 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 3 × 3 grid, with the central position of the grid being empty. The 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, participants had 7 s to move a mouse cursor to select any object that had not been selected on a previous trial (thus, all responses were correct on the first step). Once a selection was made, a white outline was displayed around the selected object until 9 s had elapsed from the start of the step. At this point, the objects in the display were pseudorandomly 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 7 s, a white outline was displayed for 2 s 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 7 s 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 s. The stimuli measured up to 7.15º of visual angle horizontally and vertically, but most of the objects were slightly smaller than that in one direction. The grid was arranged such that 0.36º of visual angle separated the outermost edges 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 from three to 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, 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 that participants were given to read the sentences was determined by the mean time that each individual participant took to read the sentences during the practice session, plus 2.5 *SD*s (see Unsworth et al., 2005, for a justification of the equivalent procedure for the operation span task).

Estimation of WM capacity

Change detection task A large body of work has employed a simple model of WM to estimate capacity from change detection tasks, based upon 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 participant detecting the change is presumed to be k/N, where k is the WM capacity and N is the number of items in the display, and $k \leq N$. Thus, the change is *not* detected by the participant with probability $(N - k)/N$, but in these cases, the participant may still correctly guess that a change occurred with probability g, even though he or she was unable to make a correct determination from memory. Thus, the probability of the participant making a hit (in signal detection terms) is $H = k/N + g(N - k)/N$. Conversely, the probability of a correct rejection on trials in which no change occurs is $CR = k/N + (1 - g)(N - k)/N$. 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, 2001, 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 was used in the present study).

This simple form of the model, however, has two important problems. First, the model predicts perfect performance whenever $N \leq k$, 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 et al. (2008) dealt with the first issue by expanding the model to include an attention parameter. That is, they assumed that on some trials participants would fail to attend to the target stimuli, and as a result, no items would be encoded into WM. Thus, given the attention parameter a and the probability $d = \min(k/N, 1)$ that the probed item is in WM, the probabilities of a hit or correct rejection in this model are

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

This expanded model can no longer be solved directly for k, because 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 et al. dealt with this issue in a way that also addresses the second criticism of the simpler model outlined above: by using maximum likelihood estimation (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 WM task An approach similar to that described for the change detection task can be applied to the SOT in order to obtain capacity estimates; however, a substantially 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 that 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 ith item has been previously selected and $C_i = 0$ when it has not, the probability that the participant will select the ith 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, since it is always equal to 1 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 $S - 1$ stimuli have previously been selected 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 $S - 1 \leq k$, $m = S - 1$, so the probability of an error becomes 0. Thus, as in the change detection task, if a participant makes an error at an early step

[1] Although Rouder et al. (2008) did not spell out the MLE procedure used, the approach taken here was to treat each trial as a Bernoulli trial with a probability determined by the formulae above and to use a brute-force search of possible values for all three parameters.

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, whereas 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 those for change detection and SOT tasks, given the complex nature of the task and how behavioral responses are made on each trial. First, assuming that the participant is fully attending to the trial, the probability that any given item on a trial will be maintained in WM in a fixed-capacity model is the same as for change detection; that is, $d = \min(k/N, 1)$. However, in the case in which an item is not maintained in WM, the participant may either guess with (unknown) probability b from one of the 12 items in the display, or the participant may elect to make no response (i.e., by responding with a "blank") with probability $1 - b$. Assuming that the participant attempts to guess, he or she has probability $g_a = 1/(12 - c)$ of guessing correctly, where c is the number of items that can be eliminated from the full range of possibilities on the basis of either items that are remembered or the number of items that have already been 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 the maximum, c is somewhat more complex, since 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 $R = 0$, $c_{\min} = c_{\max} = \min(k, N)$. Although c itself is unknown on any given trial, under the assumption that the items that 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, $E(c) = (c_{\min} + c_{\max})/2$. Consequently, we used 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

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

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 the latter case, they may again guess amongst the remaining items with probability b (for simplicity, we assume that 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 that they guess, they have probability $g = 1/(12 - R)$ of guessing correctly, under the assumption that the contents of WM are momentarily not available to them, 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

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

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

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

With these formulae, the probabilities of a correct response, incorrect response, and 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—i.e., 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) would provide no additional information over the capacity model, using a log-likelihood goodness-of-fit test on each individual participant's data.

Furthermore, to test whether the capacity models provide reliable estimates of capacity, we calculated capacity estimates on the two halves of each participant's 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 participant for each reliability estimate).

Correlation analyses

Initial examination of the data revealed that scores from the SOT task were decidedly nonnormal (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 nonparametric bootstrapping method to obtain confidence intervals (CIs) 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 the capacity estimates for each of the tasks was not simply due to an inappropriate capacity model for one or more of the tasks, we also computed pairwise correlations of accuracy (as percentages 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 might 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; see, 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 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 on SAT scores. This resulted in 15 observed variables (seven for Steps 2–8 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 nonnormal (i.e., 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 the 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 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 the participants in the SOT, 77 % of the participants in the change detection task, and 89 % of the participants in the complex span task. Reliability estimates, however, were poor for both the change detection and complex span tasks, at .32 and .45, respectively, whereas the SOT fared much better, at .77. Because the maximum population correlation between two variables is the square root of the product of their reliabilities, this implied that the maximum correlations that could be observed between these tasks were quite modest (in the range from .38 to .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 means (± SDs) of the values of a for each of the three tasks were .955 (.035) for the SOT, .857 (.127) for the change detection task, and .943 (.080) for the complex span task. This value of a for the change detection task is unusually low, since 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., one or two), which have been used in other studies and help to constrain the value of a by providing a set

[2] Although Efron and Tibshirani (1993) did not provide a means of directly calculating p values for BC_a, 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.

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 the change detection task (r = .29, p = .012) and showed a trend-level correlation for the change detection and complex span tasks (r = .18, p = .074), but they were not correlated between the SOT and the complex span task (r = .10, p = .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

The performance data (as percentages correct) at each load are displayed in Fig. 1 for all three tasks. The mean (± SD) values of k were 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; the capacity estimates for all three tasks were all significantly different from each other (all ps < .001).

Correlation analyses

The correlation between the capacity estimates in the SOT and the change detection task was .28 (p < .001), with a 95 % CI of .11–.47. The correlation between the capacity estimates in the SOT and the complex span task was –.15 (p = .086), with a 95 % CI of –.31 to .02. Finally, the correlation between the capacity estimates in the change detection and complex span tasks was –.05 (p = .600), with a 95 % CI of –.23 to .14.

The correlation between SOT and SAT was .08 (p = .464), with a 95 % CI of –.16 to .30; that between change detection and SAT was .01 (p = .954), with a 95 % CI of –.28 to .25; and that between complex span and SAT was .12 (p = .345), with a 95 % CI of –.15 to .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 a false discovery rate (FDR) correction for multiple comparisons (Benjamini & Hochberg, 1995), and only two achieved significance at an uncorrected α level of p < .05. These were the correlation between performance at Step 6 of the SOT and a three-item load in the complex span task (r = .25) and the correlation between performance on Step 8 of the SOT and a four-item load in the complex span task (r = .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 ten correlations between the loads of the change detection and complex span tasks were similar, with no correlations passing FDR correction, and only one correlation being significant without correction: the correlation between performance on eight-item sets in the change detection task and three-item sets on the complex span task (r = .20). In contrast, of the 14 correlations between performance at the various loads of the SOT and the change detection task, ten of them passed FDR correction (rs ranged from .25 to .50). The four correlations not passing FDR correction (or an uncorrected α of .05) were the correlations between Steps 2–5 of the SOT and the eight-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.

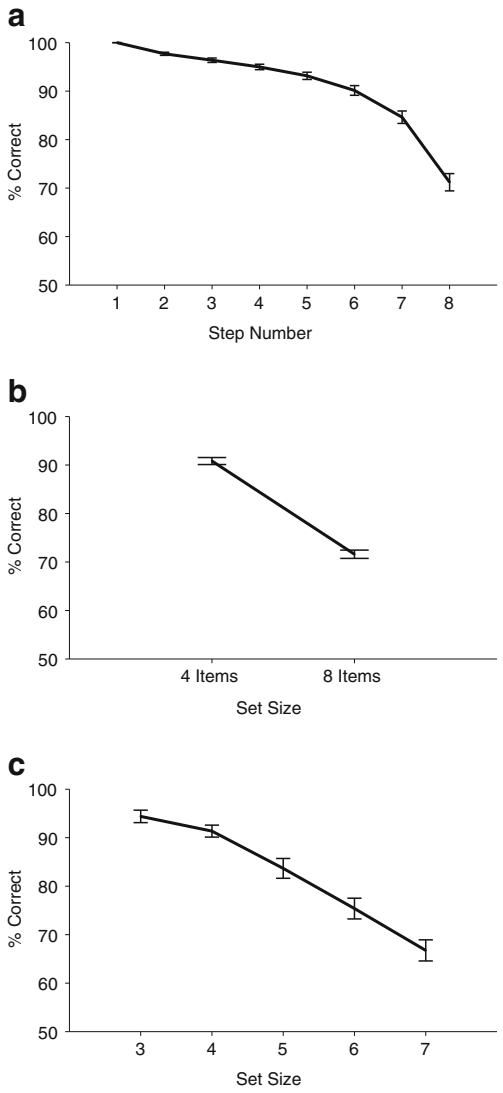

Fig. 1 Behavioral performance on the **a** self-ordering working memory task, **b** change detection task, and **c** complex span task. Error bars reflect ±1 standard error

Exploratory factor analysis

The PA factor retention decision resulted in two factors; the factor loadings are shown in a biplot in Fig. 2. All of the SOT accuracy measures and the accuracy on four-item displays of the change detection task significantly loaded on Factor 1 (all ps < .05), whereas the complex span accuracy measures for five to seven items significantly loaded on Factor 2 (all ps < .05), and the complex span accuracy measures for three and four items showed a trend toward significant loadings on Factor 2 (ps = .057 and .092, respectively). Factor loadings, confidence intervals, and p values are presented for all of the variables in Table 1. It is also noteworthy that eight-item change detection task accuracy and SAT scores approached marginal significance (p < .2) for Factors 1 and 2, respectively; as we noted in the Method section, the 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 ps < .01).

Discussion

The results of both the pairwise correlations between tasks and the EFA strongly suggest that at least two distinct cognitive abilities underlie capacity estimates and performance on the three WM tasks studied here; on the basis of 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)—whereas capacity on the SOT and change detection tasks is driven by a limit on the number of items that individuals can hold in the focus of attention at the same time (Cowan, 2001). Thus, WM and WM capacity are not necessarily unified concepts as they have been used in the literature, and depending on the task under discussion, they 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 had not previously been 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 (2001), substantially better reliability was achieved (Van Snellenberg, 2012), and likely would have been even higher with the inclusion of low-set-size trials. The results of Van Snellenberg's analyses without an attention parameter were also broadly consistent

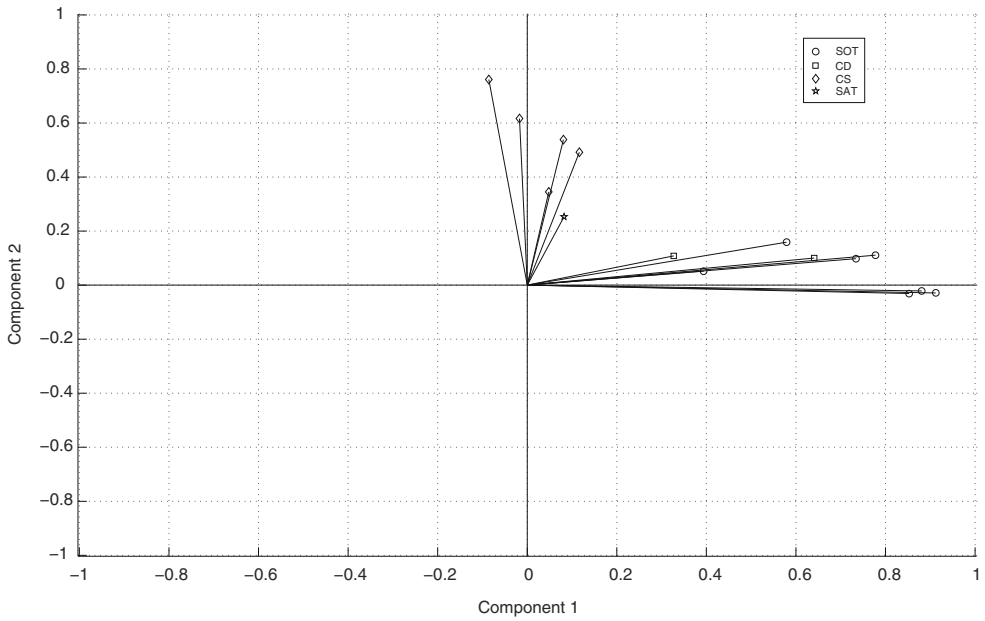

Fig. 2 Biplot of factor loadings from the exploratory factor analysis. SOT, self-ordering working memory task; CD, change detection task; CS, complex span task; SAT, SAT score

Table 1 Exploratory factor analysis loadings

	Factor 1			Factor 2		
Variable	Factor Loading	95 % CI	*p*	Factor Loading	95 % CI	*p*
SOT Step 2	.39	.03 to .65	.040	–.05	–.65 to .35	.649
SOT Step 3	.58	.17 to .88	.007	–.16	–.64 to .23	.262
SOT Step 4	.73	.42 to .91	<.001	–.10	–.41 to .11	.217
SOT Step 5	.78	.41 to .94	<.001	–.11	–.46 to .21	.222
SOT Step 6	.88	.72 to .96	<.001	.02	–.18 to .26	.994
SOT Step 7	.91	.71 to 1.02	<.001	.03	–.25 to .29	.955
SOT Step 8	.85	.56 to .95	<.001	.03	–.20 to .47	.749
CD 4-Item	.64	.06 to .87	.031	.10	–.33 to .84	.769
CD 8-Item	.33	–.11 to .67	.136	.11	–.51 to .75	.762
CS 3-Item	.12	–.14 to .56	.399	.49	–.02 to .99	.057
CS 4-Item	.05	–.13 to .36	.641	.34	–.07 to .77	.092
CS 5-Item	.08	–.16 to .40	.458	.54	.04 to 1.02	.038
CS 6-Item	–.09	–.31 to .23	.416	.76	.24 to 1.13	.011
CS 7-Item	–.02	–.26 to .19	.658	.62	.10 to 1.04	.034
SAT	.08	–.26 to .39	.528	.25	–.17 to .57	.181

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

with those reported here, 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 has occurred—have been shown to have greater reliability than change detection tasks (Johnson et al., 2013; Kyllingsbæk & Bundesen, 2009), and so may be preferable to change detection tasks in future work. Our 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 the reliability with 45 trials would be .71, much stronger than the .45 that we observed with 15 trials.

It is also worth noting that the capacity estimates for the SOT, although they were fitted reasonably well and were fairly reliable, were much higher than would be expected for a true estimate of the number of items that individuals can hold in the focus of attention (i.e., 6.2 items, as compared to 4.4 for the change detection task). Although this may to some extent be due to our sample characteristics (predominantly Columbia University undergraduates), it is almost certain that additional cognitive strategies can be brought to bear to subserve performance on this task; indeed, other work with the task has strongly suggested that this is the case (Van Snellenberg, Slifstein, et al., 2013). Nonetheless, the modest correlation that we observed between SOT and change detection capacity indicates that performance on the SOT is constrained at least 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 was entirely unrelated to the number of items that individuals can maintain in the change detection task or SOT. 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 by 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, the estimates of capacity in each of our tasks were quite high (all >4). Consequently, it is entirely possible that the strength of the correlations between tasks was limited by restriction of the range in the general cognitive abilities of our participants, or even that a relatively high-performing sample such as ours might 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, the 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. Although it is unlikely that a variable such as fatigue or motivation could produce 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 for 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 the 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. Whereas 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 the 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. Although the pattern of results that 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.

Author note E.E.S. (deceased August 17, 2012) was heavily involved throughout the preliminary and intermediate phases of the study reported in this article. 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. This work was supported by NIMH Grant No. 5P50 MH086404. The authors 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. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57,* 289–300.

Chein, J. M., Moore, A. B., & Conway, A. R. A. (2011). Domain-general mechanisms of complex working memory span. *NeuroImage, 54,* 550–559. doi:10.1016/j.neuroimage.2010.07.067

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

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Pyschonomic Bulletin & Review, 12,* 769–786. doi:10.3758/BF03196772

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–114. doi:10.1017/S0140525X01003922. disc. 114–185.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51,* 42–100. doi:10.1016/j.cogpsych.2004.12.001

Curtis, C. E., Zald, D. H., & Pardo, J. V. (2000). Organization of working memory within the human prefrontal cortex: A PET study of self-ordered object working memory. *Neuropsychologia, 38,* 1503–1510.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall/CRC.

Gibson, B., Wasserman, E., & Luck, S. J. (2011). Qualitative similarities in the visual short-term memory of pigeons and people. *Psychonomic Bulletin & Review, 18,* 979–984. doi:10.3758/s13423-011-0132-7

Gold, J. M., Fuller, R. L., Robinson, B. M., McMahon, R. P., Braun, E. L., & Luck, S. J. (2006). Intact attentional control of working memory encoding in schizophrenia. *Journal of Abnormal Psychology, 115,* 658–673. doi:10.1037/0021-843X.115.4.658

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7,* 191–205.

Johnson, M. K., McMahon, R. P., Robinson, B. M., Harvey, A. N., Britta, H., Leonard, C. J., . . . Gold, J. M. (2013). The relationship between working memory capacity and broad measures of cognitive ability in healthy adults and people with schizophrenia. *Neuropsychology, 27*, 220–229.

Kyllingsbæk, S., & Bundesen, C. (2009). Changing change detection: Improving the reliability of measures of visual short-term memory capacity. *Psychonomic Bulletin & Review, 16,* 1000–1010. doi:10.3758/PBR.16.6.1000

Lin, P.–. H., & Luck, S. J. (2012). Proactive interference does not meaningfully distort visual working memory capacity estimates in the canonical change detection task. *Frontiers in Psychology, 3,* 42.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281. doi:10.1038/36846

Osaka, N., Osaka, M., Kondo, H., Morishita, M., Fukuyama, H., & Shibasaki, H. (2004). Theneural basis of executive function in working memory: An fMRI study based on individual differences. *NeuroImage, 21,* 623–631. doi:10.1016/j.neuroimage.2003.09.069

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics, 44,* 369–378. doi:10.3758/BF03210419

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences, 105,* 5975–5979. doi:10.1073/pnas.0711295105

Shipstead, Z., Redick, T. S., Hicks, K. L., & Engle, R. W. (2012). The scope and control of attention as separate aspects of working memory. *Memory, 20,* 608–628. doi:10.1080/09658211.2012.691519

Smith, E. E., Geva, A., Jonides, J., Miller, P., Reuter-Lorenz, P., & Koeppe, R. A. (2001). The neural basis of task-switching in working memory: Effects of performance and aging. *Proceedings of the National Academy of Sciences, 98,* 2095–2100.

Smith, E. E., & Van Snellenberg, J. X. (2011). Capacity and processing deficits in working memory in schizophrenia. *Schizophrenia Bulletin, 37,* 228.

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature, 428,* 751–754. doi:10.1038/nature02466

Todd, J. J., & Marois, R. (2005). Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cognitive, Affective, & Behavioral Neuroscience, 5,* 144–155. doi:10.3758/CABN.5.2.144

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114,* 104–132. doi:10.1037/0033-295X.114.1.104

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. doi:10.3758/BF03192720

Van Snellenberg, J. X. (2012). *An investigation of the neural correlates of working memory in healthy individuals and individuals with schizophrenia*. PhD dissertation, Columbia University. Retrieved from http://hdl.handle.net/10022/AC:P:13157

Van Snellenberg, J. X., Girgis, R. R., Read, C., Thompson, J. L., Weber, J., Wager, T. D., … Smith, E. E. (2013). Individuals with schizophrenia fail to show normative inverted-U activation in response to fine-grained working memory load manipulation [Abstract]. *Schizophrenia Bulletin*, *39*, S251–S252.

Van Snellenberg, J. X., Slifstein, M., Read, C., Weber, J., Thompson, J. L., Wager, T. D., … Smith, E. E. (2013). *Dynamic shifts in brain network activation during working memory task performance*. Manuscript submitted for publication.

Van Snellenberg, J. X., Wager, T. D., Abi-Dargham, A., Urban, N., & Smith, E. E. (2010). Parametric variation in working memory demand in patients with schizophrenia: A behavioral and neuroimaging pilot study [Abstract]. *Biological Psychiatry, 67,* 158S.

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature, 428,* 748–751. doi:10.1038/nature02447

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature, 440,* 91–95. doi:10.1038/nature04262