Published in final edited form as:

Vis cogn. 2016 ; 24(9-10): 435–446. doi:10.1080/13506285.2017.1301613.

Assessing the evidence for a cue-induced trade-off between capacity and precision in visual working memory using mixture modelling and Bayesian model comparison

Andrea Bocincova, Amanda E. van Lamsweerde, and Jeffrey S. Johnson

Department of Psychology and Center for Visual and Cognitive Neuroscience, North Dakota State University

Abstract

There is considerable debate regarding the ability to trade mnemonic precision for capacity in working memory (WM), with some studies reporting evidence consistent with such a trade-off and others suggesting it may not be possible. The majority of studies addressing this question have utilized a standard approach to analyzing continuous recall data in which individual-subject data from each experimental condition is fitted with a probabilistic model of choice. Estimated parameter values related to different aspects of WM (e.g., the capacity and precision of stored items) are then compared using statistical tests to determine the presence of hypothesized differences between experimental conditions. However, recent research has suggested that the standard approach is flawed in several respects. In this study, we adapted the methods of Roggeman et al. (2014) and analyzed the data using the standard analytical approach and a more rigorous Bayesian model comparison (BMC) approach. The second approach involved generating a set of probabilistic models whose priors reflect different hypotheses regarding the effect of our key experimental manipulations on behavior. Our results demonstrate that these two approaches can produce notably different results. More specifically, the standard analysis revealed that a high-versus a low-load cue resulted in higher capacity and lower precision parameter estimates, suggesting the presence of a trade-off between capacity and precision. However, the more rigorous BMC analysis revealed that it was very unlikely that participants employed a behavioral strategy in which they sacrificed mnemonic precision to achieve higher storage capacity. In light of these differences, we advocate for a more stringent approach to model selection and hypothesis testing in studies implementing mixture modeling.

Keywords

color working memory; capacity; precision; trade-off; pre-cues; Bayesian model comparison

Working memory (WM) refers to our ability to store and manipulate information over short time periods when that information is no longer present in the environment (Baddeley & Hitch, 1974). Behavioral tests of WM have shown that performance remains quite good

Andrea Bocincova (Corresponding Author), North Dakota State University, Dept. 2765, PO Box 6050, Fargo, ND 58108, 701-231-8934, andrea.bocincova@ndsu.edu.

The authors declare no conflict of interest.

when the number of items to be stored is small (e.g., ~3–4 simple objects), but declines rather dramatically as storage demands are increased (Cowan, 2001; Luck & Vogel, 1997). Despite these limits, humans are capable of successfully managing information-rich environments by selectively attending to and remembering only the subset of information most relevant to their current goals. For example, studies have shown that people can selectively store behaviorally relevant information while ignoring task-irrelevant information, and that this ability varies between individuals and is predictive of WM capacity (Vogel, McCollough, & Machizawa, 2005). Empirical evidence also suggests that the precision of WM representations declines as the number of items stored increases (Palmer, 1990; Wilken & Ma, 2004; Bays & Husain, 2008; Zhang & Luck, 2008; Bays, Catalao, & Husain, 2009). Selecting a subset of information for storage, therefore, may also serve as a useful strategy when information needs to be remembered more precisely (Donkin, Kary, Tahir, & Taylor, 2016; Klyszejko, Rahmati, & Curtis, 2014).

Although our ability to limit access to WM resources is well established, it remains unclear whether it is possible to increase the amount of information that can be remembered by storing items in WM with lower precision. Answering this question is relevant to ongoing debates regarding the nature of WM representations, with two prominent views of WM making contrasting predictions. On the one hand, flexible resource approaches state that WM resources can be allocated to a small number of items that are stored with high precision, or distributed across a large number of items that are stored with low precision (Palmer, 1990; Wilken & Ma, 2004; Bays & Husain, 2008; Zhang & Luck, 2008; Bays et al., 2009). On the other hand, the discrete slots view holds that there is a fixed upper capacity limit that cannot be exceeded by storing additional items with lower precision (Zhang & Luck, 2008; 2011). Despite its relevance, the issue has proved difficult to resolve. Most studies have examined this question using variants of the cued recall task depicted in Figure 1, which requires participants to remember a small number of memory targets across a brief, unfilled delay interval, and to select a specific cued target from a continuous representation of the feature space at test. The resulting recall distributions can then be fit with probabilistic models that make it possible to derive separate estimates of the number of items that were stored and their precision (see, e.g., Bays et al., 2009; Zhang & Luck, 2008, and discussion below). Using these methods, Zhang & Luck (2011) found that observers were unable to increase the number of items that were stored by storing items at lower precision, even when provided with strong behavioral incentives to do so. Similarly, Murray, Nobre, Astle, and Stokes (2012) observed no capacity-precision trade-off when payoff incentives in the task were manipulated to emphasize either the quality or quantity of memory. Another study showed that precision could be willfully controlled, but only when the overall number of stored items was below an individual's capacity (Machizawa, Goh, & Driver, 2012). Taken together, these studies suggest that it may not be possible to arbitrarily increase the number of items stored by representing items with lower precision.

In contrast to these findings, Roggeman, Klingberg, Feenstra, Compte, and Almeida (2014) reported results of a study of visuospatial WM that revealed an apparent ability to increase capacity at the expense of precision. In their study, pre-cues were presented to bias participants' expectations regarding the likely number of to-be-remembered items that would be presented on that trial. On the basis of a biologically constrained model of

visuospatial WM (Edin et al., 2009), they predicted that cues leading participants to expect a large number of items on a given trial (5 versus 3) would engage areas of the frontal cortex involved in implementing executive control functions. Once engaged, these control areas send top-down excitatory input to WM storage sites in the parietal cortex. In the model, such inputs have the effect of increasing the overall level of activation within posterior storage networks, which allows greater numbers of items to be stored. However, such ‘boost’ inputs also increase non-specific background noise and the potential for interference between stored items, which produces an associated reduction in mnemonic precision. Their results generally supported this prediction: although analyses based on individual subject data fits produced somewhat mixed results, analysis of capacity and precision parameters that were estimated from recall data and aggregated across participants revealed that capacity was higher and precision was lower when pre-cues led participants to expect a large number of items on a given trial. Additionally, correlation analysis assessing the relationship between cue-related changes in capacity and precision revealed a robust relationship between cue-related increases in capacity and decreases in precision at the individual subject level. Based on a closer inspection of these correlations, the authors proposed that the relatively small group-level results were likely driven by a subset of participants who were able to successfully utilize the cue to boost their capacity at the expense of precision.

Each of the studies reported thus far utilized what has become the standard analytical approach in this area, in which individual-subject data from each experimental condition is fitted with a probabilistic model of choice and parameter estimates are derived for each participant and condition. For example, the standard mixture model of Zhang and Luck (2008) produces three different parameters of interest: the mean and standard deviation of the recall distribution, assumed to reflect average mnemonic accuracy and resolution, respectively, when the cued item is successfully stored in WM, and the guess rate (g), assumed to reflect the likelihood that recall responses were selected at random from the feature space when the item was not successfully stored. The guess rate is often used to derive an estimate of the likelihood of successful storage ($1-g$), which is used as a proxy for an individual’s WM capacity. Once these parameters are estimated, differences in group means across conditions are then assessed using common null-hypothesis statistical tests (e.g., ANOVA or t-tests). For example, in the study of Roggeman et al. (2014), evidence for a capacity-precision trade-off was obtained by running separate ANOVAs on two parameters of the mixture model (the guess rate and the standard deviation) proposed to reflect the capacity and precision of WM, respectively.

The use of these methods has increased dramatically in recent years, owing in part to the availability of easy-to-use, open-source software tools for fitting data with a variety of different models (Suchow, Brady, Fougner, & Alvarez, 2013; memtoolbox.org). Along with this increase in ease of application, however, has come a growing awareness of the challenges of applying these seemingly straightforward methods and drawing conclusions based on their results (Kruschke, 2010; Oberfeld & Franke, 2013; Suchow et al., 2013). The first challenge is in selecting an appropriate model to fit the data. In addition to the standard mixture model, numerous additional models have been proposed that embody different assumptions regarding the nature of storage in WM, and comparisons of the goodness of fit of these various models has been used as a means of adjudicating between different theories

of WM. In their attempt to provide a more comprehensive comparison of the currently documented model families, van den Berg, Awh, and Ma (2014) reported a factorial comparison of 32 different variants of the mixture model that embodied different assumptions regarding whether memory resources are discrete or continuous, whether WM capacity has a fundamental limit or not, and the extent to which recall errors may be due to mis-remembering the spatial location of remembered features (as proposed by Bays et al., 2009). Results revealed strong evidence against all six of the models represented in the literature at the time, highlighting the difficulty of selecting an appropriate model and the need to include and test more complete sets of models in order to address questions pertaining to WM storage.

Although the use of model comparison techniques is becoming more widespread in this area, it is still common to select a model using less formal criteria. For example, a particular model might be selected not because it provides the best possible fit to the data, but because it matches the models that have been used to address similar problems in the past, facilitating the comparison of results with previous relevant findings. Although model choice carries with it certain assumptions regarding the structure of WM, models are often used in an almost theory-neutral fashion to assess the effect of a given experimental manipulation on different aspects of performance. However, even if the validity of this approach is accepted, it is not without its limitations. Chief among these is the fact that analysis of parameter estimates using separate ANOVAs does not account for between-subjects variations in the reliability of each estimate or take into account naturally existing co-variations between the parameter estimates themselves (Kruschke, 2010; Oberfeld & Franke, 2013; Suchow et al., 2013). More specifically, fitting the model separately to each experimental condition assumes that the parameter distributions of the different conditions vary independently, and the following statistical analysis cannot rule out whether a model with a different set of assumptions would have provided a better description of the data. Moreover, the use of an identical model to fit the data across all subjects makes it difficult to identify the unique strategies that different participants may have adopted during the task; for example, whether a subset of participants in the Roggeman et al. (2014) study utilized the high-load cue to increase capacity at the expense of precision, as the authors propose, or what alternate strategies, if any, may have been adopted by the other study participants.

Recently, Dowd, Kiyonaga, Beck, and Egner (2015) introduced a hypothesis-driven variant of the factorial model comparison approach developed by van den Berg, Awh, and Ma (2014) that overcomes many of the limitations noted above. This approach involves developing a set of models whose priors explicitly incorporate different hypotheses regarding the expected effect of a given experimental manipulation on the different parameters of the model. These models are then fitted to each participant's data and the best fitting model is identified for each individual as well as for the whole sample using Bayesian model comparison methods (Stephan, Penny, Daunizeau, Moran, & Friston, 2009). Instead of inferring the presence of the hypothesized effects of different experimental manipulations from a statistical analysis of each estimated parameter separately, this approach selects the model corresponding to an experimental hypothesis that best describes the experimental data as a whole. Moreover, the possibility that different participant's data will be better fit by one

or the other model makes it possible to assess the various strategies participants may have adopted across conditions.

The primary goals of the present study were to affect a more stringent test of the trade-off hypothesis, and to compare the outcomes obtained using the standard approach and the more rigorous model-comparison approach described above. To do this, participants completed a color recall WM task adapted from the study of Roggeman et al. (2014), in which pre-cues were used to manipulate trial-by-trial expectations about the upcoming behavioral demands of the task. According to Roggeman et al., pre-cues leading participants to expect that a large number of items will need to be remembered on a particular trial should lead to selective engagement of a top-down boost signal that increases activity in storage-related regions of the parietal cortex. This should lead to an increase in the likelihood of storing the cued item in WM on trials with high load demand, but at the cost of decreased precision. In order to evaluate the evidence for such a trade-off, and to explore how the choice of analysis method may impact observed outcomes, we implemented two different approaches to analyzing our data. First, the data was analyzed using the standard analytic approach in which a probabilistic model, the standard mixture model of Zhang and Luck (2008), was fitted to each individual's data separately for each experimental condition, and estimates of model parameters thought to index the precision and capacity of WM were obtained. Two separate ANOVAs were then conducted to assess whether the type of cue had systematic effects on either parameter. Secondly, a more computationally intensive approach was adopted in which a set of hypothesis-driven probabilistic models was created that explicitly incorporated different predictions about the expected effects of our experimental manipulations on parameters of the mixture model. Bayesian model comparison (Dowd et al., 2015; Stephan et al., 2009) was then used to formally assess the ability of each model to fit the data.

Observed results provided support for diametrically opposite conclusions depending on the method of analysis. Whereas the standard approach generally supported the existence of the proposed capacity-precision trade-off, replicating the findings of Roggeman et al. (2014), the hypothesis-driven Bayesian model comparison approach revealed that the best fitting model, both at the individual-subject and group levels, was one that assumed equal capacity and precision across conditions. These opposing outcomes highlight the challenges of drawing inferences from recall data using currently available tools, and the utility of more formal model comparison approaches.

Methods

Participants.

Sixty undergraduate students (33 females) 18–33 years ($M = 19.51$, $SD = 2.61$) from North Dakota State University participated in this study for course credit¹. All participants reported normal or corrected to normal vision and no prior or current neurological or mental disorders. Each participant provided written informed consent. All experimental protocols were approved by the North Dakota State University IRB.

¹Thirty participants had concurrent EEG recorded while performing the behavioral task. The EEG data will not be reported here.

Materials and procedure.

Stimulus presentation and response recording was controlled by a PC running Matlab (Mathworks, Inc.) with Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). Sample displays consisted of 2 or 4 colored squares subtending $1.5^\circ \times 1.5^\circ$ of visual angle. Individual colors were selected at random from a set of 180 colors equally distributed in CIELAB (1976) color space (centered at CIE $L^*a^*b^*$ coordinates: $L=70$, $A=28$, $B=12$). Stimuli were presented against a light grey background on the surface of a 19" cathode ray tube monitor with a refresh rate of 100 Hz, at a viewing distance of 60 cm. Stimuli were presented at unique locations randomly chosen from a set of eight possible positions (four on each side of the screen) within two 4×2 square grids on either side of fixation, with the constraint that an equal number of stimuli were presented on each side of the screen. The test display contained white squares at the sample stimuli locations: a filled square at a randomly selected target location and outlines at the non-target locations. These were surrounded by a continuous color wheel, 15.28° in radius, centered at fixation. The color wheel contained each of the possible sample colors equally distributed in 2° steps. The orientation of the color wheel was randomly rotated across trials.

The sequence of events in a typical trial can be seen in Figure 1. Each trial started with the appearance of a fixation cross that remained visible at the center of the screen throughout the trial. After 1000 ms, a cue (75% valid) was displayed above fixation for 500 ms, indicating whether 2 (low-load cue; Cue2) or 4 (high-load cue; Cue4) sample items would be presented on that trial. The cue was followed by a 1500 ms cue-sample interval (± 500 ms jitter) and the 150 ms presentation of a sample display containing either 2 (set size 2; SS2) or 4 (set size 4; SS4) colored squares. Participants were instructed to remember the color of all the sample display items across a 1000 ms blank delay. Following the delay, participants were asked to report the color of the cued target by moving a set of crosshairs around the color wheel and clicking on the remembered color using a computer mouse. Once the participant moved the mouse, the center of the filled white square changed continuously to match the currently selected color. After the response was made, the correct color was shown as a thick border around the cued item. A red bar also marked the position of the correct color on the color wheel. Instructions stressed the importance of accuracy and informed the participants that responses were not timed. The experiment consisted of 480 trials total: 360 trials in which the pre-cue matched the number of items presented on that trial (Valid Cue trials; 160 Cue2/SS2 and 160 Cue4/SS4), and 120 trials in which it did not (Invalid Cue trials; 60 Cue2/SS4 and 60 Cue4/SS2), randomly intermixed. The entire session was completed in a single session lasting ~1.5 hours, and was broken up into 10 blocks of 48 trials separated by short breaks.

Data Analysis

Mixture modeling.—Participants' data were analyzed using the MemToolbox (Suchow, Brady, Fougine, & Alvarez, 2013; memtoolbox.org) and custom scripts written in Matlab. Response errors were calculated by subtracting the actual target color value (in degrees) from the reported color value on a given trial. To separately examine the amount of information stored and the precision of the stored information, the resulting response error distributions were then modeled as a mixture of a von Mises (the circular analog of a normal

distribution) and a uniform distribution using maximum likelihood estimation (Zhang & Luck, 2008). The mean (μ) and standard deviation (s.d.) of the von Mises distribution can be used to estimate the accuracy and precision (1/s.d), respectively, of representations that were successfully stored in WM. The uniform distribution, by contrast, can be used to estimate the probability, g , that recall responses were selected at random from the color wheel. This value can then be used to estimate the overall probability, $1-g$, that the probed item was present in memory at test (i.e., of memory capacity, denoted as P_m). Although several different models have been proposed to account for performance in recall WM tasks, the two-component model of Zhang and Luck (2008) was chosen to allow a direct comparison of the present results with those of Roggeman et al. (2014).

Standard statistical analysis of model parameter estimates.—For our first set of analyses, we adopted the standard approach described above, in which, for each participant, separate estimates of the mixture model parameters (in particular, g and s.d.) were obtained for each combination of set size and cue type, and separate repeated-measures ANOVAs were used to assess differences between conditions for each parameter. Specifically, to determine whether the capacity and precision of WM differed as a function of Cue Type (Cue2 vs. Cue4), obtained capacity and precision estimates were analyzed with separate 2 (Set Size: 2, 4) $\times$ 2 (Cue Type: Cue4, Cue2) within-subjects ANOVAs. On the basis of the hypothesized trade-off between capacity and precision resulting from the cue manipulation, we expected P_m ($1-g$) to be greater on high-load cue (Cue4) versus low-load cue (Cue2) trials, but that this would be achieved at the expense of lower precision. In other words, for each 2-way ANOVA, we expected to observe a significant main effect of cue type on the relevant parameter, and possibly an interaction between Cue Type and Set Size. Most importantly, following the logic of Roggeman, et al. (2014), we expected to observe a significant correlation between the cue-related change in capacity and the cue-related change in precision across participants. That is, participants exhibiting a large cue-related increase in capacity should show a correspondingly large cue-related decrease in precision. Further, we expected that this pattern would be observed even in the absence of significant group-level effects in the ANOVA, which might occur if only a subset of the participants were capable of using the cue to boost capacity in the manner proposed (for similar arguments, see Roggeman, et al.). To assess this possibility, Pearson r was calculated to measure the degree of correlation between observed cue-related changes in capacity (Cue4-Cue2) and precision for each participant.

Hypothesis-driven probabilistic models and Bayesian model comparison.—

The second set of analyses involved generating a set of hypothesis-driven models that correspond to the possible effects of the key experimental manipulation, and selecting the model that best describes the data. Specifically, we generated four different models that embody different assumptions regarding the expected effect of our cueing manipulation on obtained recall estimates. In each model, the corresponding hypothesis is expressed as a change in the prior probability distributions, or “priors”, that are applied to the estimates of both s.d. and g for each model. The first model (Model 0) we implemented corresponds to the standard mixture model with priors reflecting the assumption that recall responses across conditions are drawn from distributions with identical g and s.d.; that is, that the cueing

manipulation has no effect on either s.d. or g across conditions. The second model (Model 1) represents a mixture model in which both s.d. and g are free to vary independently across experimental conditions. The parameter estimates obtained from this model would be comparable to those expected if model parameters were estimated separately for each condition, as in the standard analysis described above, but with no specific hypothesis regarding how these parameters will differ across conditions. The third model (Model 2) implements the trade-off hypothesis in which a high-load cue is expected to increase capacity ($g_{\text{Cue4}} < g_{\text{Cue2}}$) at the expense of decreased precision ($s.d._{\text{Cue4}} > s.d._{\text{Cue2}}$). The final model (Model 3) implements the opposite form of a trade-off, in which a low-load cue leads to reduced capacity ($g_{\text{Cue4}} > g_{\text{Cue2}}$) with the consequence that each item is stored with higher precision ($s.d._{\text{Cue4}} < s.d._{\text{Cue2}}$). This would be expected to occur if the low-load cue induced participants to adopt a strategy of storing a subset of the memory-display items with higher precision.

To compare the different experimental hypotheses represented by each model this approach uses model evidence, $p(D/M)$, which is the probability of obtaining the observed data (D) given a specific model (M). The model evidence, or marginal likelihood, of the data from a subject s , D^s , for a model M is computed by integrating over the parameters within the parameter vector θ_M of the model M :

$$p(D^s|M) = \int p(D^s|\theta_M, M) p(\theta_M|M) d\theta_M$$

Because the integration is in many cases computationally difficult and analytically intractable, an approximation to the model evidence, also referred to as log-evidence, was computed instead (Stephan et al., 2009). To do so, each model described above was fit to the data of individual participants separately for SS2 and SS4 and the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002), a measure of model comparison based on the goodness of fit and complexity, was calculated for each. Total log likelihoods were then computed from the obtained DIC values and submitted to a Bayesian model selection (BMS) routine implemented in SPM12 (spm_BMS; <http://www.fil.ion.ucl.ac.uk/spm/software/spm12/>). This routine results in two different outputs of interest: the exceedance probability (P^* ; values can vary between 0 and 1), which reflects the probability that the specific strategy implemented in a given model is more likely than those embodied by the other models, and the expectation of the posterior (posterior; values can vary between 0 and 1), which estimates the likelihood that the data of a randomly selected participant was generated by a given model (Stephan et al., 2009).

Bootstrap analysis.—In light of the results of the Bayesian model comparison, reported below, we also conducted a bootstrap analysis to estimate the probability of obtaining a significant correlation between the cue-related changes in capacity and precision even if no true cue-related differences between the high-load and low-load cue existed. The parameter estimates derived by the standard mixture model (specifically, g and s.d.) naturally exhibit a small negative correlation and it has been suggested that correlations between parameter estimates derived from the same set of data can produce inflated results (Suchow et al., 2013). Although the measures that are being related in this study, and in that of Roggeman et

al. (2014), are difference scores generated from two independent sets of data, these measures may still be expected to naturally covary. In order to test this possibility, each participant's recall errors were randomly re-assigned to either the high-load or low-load cue condition and estimates of g and $s.d.$ were recalculated by fitting the standard mixture model to the newly generated data. The correlation analysis described as part of the standard analysis above was then repeated using the obtained differences in g and $s.d.$ values across cuing conditions. Because such reorganization of the data eliminates any systematic relationship between cue type and response error, a significant correlation in the difference scores for capacity and precision would suggest a tendency towards the presence of a trade-off that is unrelated to the variable of cue type. This process was repeated 1000 times in order to estimate the empirical distribution of correlation values, and the likelihood of obtaining the observed correlation values in the absence of an effect of cue type.

Results

Standard statistical analysis of model parameter estimates.

The goal of this experiment was to determine whether a high-load cue increases color memory capacity at the expense of precision, as has been observed in studies of visuo-spatial WM (Roggeman et al., 2014). Differences in P_m and precision across conditions were each analyzed with a 2 (Set Size: 2, 4) $\times$ 2 (Cue Type: Cue4, Cue2) within-subjects ANOVA.

Results are depicted in Figure 2A and 2B. As observed in previous research, there was a main effect of set size for both P_m [$F(1,59) = 183.81, p < .001, \eta_p^2 = 0.76$] and precision [$F(1,59) = 98.05, p < .001, \eta_p^2 = 0.62$]. P_m was lower (guessing was more likely) at SS4 ($M = 0.75, SD = 0.14$) than at SS2 ($M = 0.95, SD = 0.06$), and precision estimates were lower (information was stored more coarsely) at SS4 ($M = 0.047, SD = 0.011$) than at SS2 ($M = 0.059, SD = 0.012$).

Of more relevance to our goals, the main effect of cue type (Cue2, Cue4) was not significant for P_m [$F(1,59) = 1.65, p = .20, \eta_p^2 = 0.03$], but was marginally significant for precision [$F(1,59) = 3.21, p = .08, \eta_p^2 = .05$] suggesting that the high-load cue may have led to lower precision compared to the low-load cue across both set sizes. Critically, there was a significant Cue Type $\times$ Set Size interaction for P_m [$F(1,59) = 4.86, p = .03, \eta_p^2 = 0.08$]. In the SS4 condition, P_m was higher when the memory display was preceded by a high-load cue versus a low-load cue ($M_{cue4} = 0.76, SD_{cue4} = 0.13; M = 0.73_{cue2}, SD_{cue2} = 0.16$; pairwise comparison marginally significant $p = .057$). There was no significant Cue Type $\times$ Set Size interaction for precision [$F(1,59) = 2.86, p = .10, \eta_p^2 = 0.05$].

Although the ANOVA results were somewhat mixed, Pearson's correlation analysis revealed a robust relationship between the P_m cue effect and the precision cue effect at both set sizes (SS2: $r = -.287, p = .026$; SS4: $r = -.576, p < .001$; see Figure 2C and 2D), as predicted. Specifically, for a subset of participants ($\sim 20/60$ at SS4; those falling in the lower right quadrant of Figure 2D) the high-load cue produced an increase in capacity (lower g) accompanied by a cue-related decrease in memory precision (increased $s.d.$), as predicted by the Roggeman et al. (2014) model.

Hypothesis-driven probabilistic models and Bayesian model comparison.

Results of the standard analysis generally support the existence of a cue-induced trade-off between capacity and precision in WM. Our second analysis affected a more rigorous test of this possibility by explicitly incorporating different assumptions into variants of the standard mixture model, and using Bayesian model comparison to determine which of several models best captures the strategies adopted by participants across the experimental conditions of interest. The relative DIC values (the difference between the DIC value of the best fitting model and all the other models) for each participant can be seen in Figure 3. In light of the findings from the standard analysis described above, we expected that, at least for a subset of participants (those falling in the lower-right quadrant of Figure 2D), the best fitting model would be Model 2, which incorporates the assumption that the high-load cue would produce an increase in capacity at the expense of precision. Contrary to the trade-off hypothesis, however, DIC revealed that a model with constant s.d. and constant g across both experimental conditions (Model 0) provided the best fit for 58 out of 60 participants at SS4 and 60 out of 60 participants at SS2. This was further confirmed by the results of the BMS, which showed that this model was superior to all other models (see Table 1). The exceedance probability for Model 0 reached $P^* > .999$ in both set sizes indicating strong evidence in favor of this model and the null hypothesis (Rigoux, Stephan, Friston, & Daunizeau, 2014). Additionally, the expected posterior probability showed a high likelihood that the data of a random participant would be generated by this model ($\text{posterior}_{SS2} = .9517$, $\text{posterior}_{SS4} = .9505$). Together, these findings invite a reconsideration of the results of the standard analysis, which suggested that the cueing manipulation induced a trade-off between capacity and precision in WM.

Bootstrap analysis.

To address the discrepancy between the results presented in the previous sections, we conducted a bootstrap analysis that allowed us to determine the likelihood of achieving a statistically significant trade-off between capacity and precision while controlling for systematic cue-related effects. The results of this analysis revealed that obtaining a statistically significant correlation between the difference scores for Pm and precision (Cue4-Cue2) was very likely; a significant correlation was found in 99.6% of the 1000 iterations that were run. To ensure that these results weren't due to the different numbers of trials in each condition (180 trials Cue4SS4 versus 60 trials Cue2SS4), we repeated the bootstrap analysis, sampling sixty random responses from the Cue4/SS4 condition to match the number of trials between the two conditions. The analysis again revealed a very high proportion of significant correlations (99.9% significant), suggesting that the observed correlations were not due to differing numbers of trials in each condition.

Discussion

The central aim of the present study was to evaluate evidence for the existence of a capacity-precision trade-off using two different approaches: a standard analytical approach utilized in the field, and a more rigorous, hypothesis-driven approach implementing Bayesian model comparison. The use of the two different approaches provides an important methodological

contribution by drawing attention to the possible spuriousness of results generated by the commonly implemented approach to analyzing continuous recall data.

Using the standard approach, in which task-related changes in mixture model parameters (g and $s.d.$) were assessed through separate ANOVAs and correlation analysis was used to assess covariation in parameters at the individual-subject level, we found results consistent with the existence of the predicted capacity-precision trade-off. Most compellingly, we observed a strong correlation between cue-related changes in capacity and precision at both set sizes tested. This analysis suggested that, in keeping with the findings of Roggeman et al. (2014), a significant minority of participants (~20/60 at SS4) could use the high-load cue to increase their WM capacity at the expense of mnemonic precision. Moreover, this result suggests that the trade-off generalizes to color WM.

Although the analytic approach adopted above represents a fairly common approach in this area, there are potential issues that make it less than ideal for assessing the hypothesis in question. For one, the choice to run separate ANOVAs on each of the mixture model parameters of interest ignores the fact that these parameters are known to be weakly negatively correlated within each participant; that is, as more responses are attributed to guessing (the uniform distribution), $s.d.$ tends to decrease and vice versa (Suchow et al., 2013). Moreover, this approach does little to inform our understanding of what may be going on with those participants exhibiting the predicted cue-related trade-off effect, versus those showing no effect, or the opposite effect (i.e., an increase in guessing and decrease in $s.d.$ in response to the low-load cue). To address these issues, we adopted a model comparison approach that made it possible to account for the relationship between g and $s.d.$ in a single model, and to formally assess the different strategies participants may have adopted in the different cuing conditions. Specifically, we implemented four different models that embodied different assumptions regarding the relationship between g and $s.d.$ across conditions. The first model assumed constant g and $s.d.$ across conditions. The second model assumed variable g and $s.d.$ across conditions, without specifying the direction of these differences. The final two models implemented different variants of the trade-off hypothesis; the first assumed that g would decrease (capacity will increase) and $s.d.$ would increase (precision will decrease) in response to the high-load cue, whereas the second assumed that g would increase and $s.d.$ would decrease in response to the low-load cue. Results of this analysis showed that the data was best captured by the first model in which g and $s.d.$ were assumed to be constant across cuing conditions. These results provide little evidence to support the existence of a strategic increase in capacity accompanied by decreased precision resulting from the cue manipulation. Instead, the analysis showed that the capacity and precision parameter estimates for the two experimental conditions that showed significant differences in the standard approach can be best described as originating from the same response distribution.

In addition to addressing the plausibility of the group level differences observed in our study, we further attempted to critically examine the likelihood of a significant correlation between cue-related effects in capacity and precision. This correlation has been taken as evidence for the existence of a capacity-precision trade-off (Roggeman et al., 2014); however, this correlation could simply be an artifact of naturally occurring covariation between the

parameters of the standard mixture model (Suchow et al., 2013). In keeping with this possibility, bootstrap analysis showed that a significant correlation was produced in over 99% of cases, despite the fact that response errors were randomly assigned to conditions.

The results of the current study raise the question of whether a similar spurious trade-off could have been observed in the Roggeman et al. (2014) study. It is important to note that any application of our conclusions to their study is limited by the design differences between the studies. These differences included the nature of the memory task (spatial versus color WM), and the recall demands (reporting the color of *one* of the items from the memory display compared to reporting the location of each item). That being said, it seems unlikely that any of these design differences would affect the hypothesized strategies the participants were using in the separate cueing conditions of our study. Considering their similarly small and variable group-level effects and the possibility of inflated correlations between cue-related effects in capacity and precision due to the model-inherent correlation of parameter estimates, the evidence for a trade-off in their study should also be treated with caution.

With respect to the ongoing debate concerning the capacity-precision trade-off, our results do not provide support for the ability of participants to strategically increase the amount of stored information at the expense of precision. However, a lack of evidence for the predicted trade-off in the present study should not necessarily be taken as unequivocal support for the discrete slots view. It remains possible that other methods of testing for this trade-off could reveal such evidence, or that other models not tested here could have provided a better fit to the data. In keeping with the former possibility, a recent study by Fougny et al. (2016) implementing the standard approach revealed that incentives to favor storage capacity over mnemonic precision did produce the expected trade-off when the standard recall task was changed so that participants had to report the colors of all of the remembered items, rather than reporting a single, randomly chosen item. This trade-off was only observed if the response type was blocked and the cue emphasizing storage capacity was presented following encoding; a cue presented at test revealed no evidence of a trade-off. This finding suggests that task specifications influence how information is encoded and maintained in WM.

Future research in this area could benefit by focusing on whether discrepant findings between studies testing for a capacity-precision trade-off are due to differences in experimental paradigms, selected models, or limitations of the chosen analytical approach. More generally, results of the present study suggest that great care should be taken when using probabilistic modeling approaches to characterize and draw conclusions from recall data. Whenever possible, we recommend adopting a more rigorous modeling approach in which specific hypotheses regarding the expected effects of different experimental manipulations are explicitly incorporated, and model comparison techniques are used to determine the best fitting model for your data (as in Dowd et al., 2015). This approach has the advantage of making it possible to simultaneously assess the impact of manipulated variables on the model's parameters, which may reduce the likelihood of drawing spurious conclusions based on naturally existing covariation between parameters. Additionally, because the goodness of fit of each model is assessed at both the individual subject and group levels, this approach has the potential to inform our understanding of the many

potential strategies observers may have adopted. Nonetheless, if the standard approach is adopted, the underlying assumptions of the chosen model should be carefully considered, and the presence of the observed differences in parameters should be evaluated using different models to avoid drawing conclusions based on model-specific experimental effects.

In conclusion, the results presented here demonstrate several potential pitfalls of the standard approach to the implementation and statistical analysis of mixture modeling data, and advocate for a more rigorous approach to mixture modeling using Bayesian model comparison techniques. To summarize, using standard methods of analyzing WM recall data, our results appeared to support the existence of a trade-off between capacity and precision in color recall, similar to those previously observed in spatial WM (Roggeman et al., 2014). However, implementation of Bayesian model comparison techniques showed that the observed trade-off was most likely an artifact of the analysis method, rather than a genuine effect of the cue on storage. Testing hypotheses by explicitly modeling the effects of experimental manipulations can help minimize some of the limitations of the standard approach and provide stronger evidence for observed effects.

Acknowledgments

We would like to thank Emma Wu Dowd for her assistance with the Bayesian Model Comparison analysis, and two anonymous reviewers for their very helpful comments and suggestions. This research was supported by the National Institute of Health under Grant R15 MH105866-01 awarded to J.S.J.

References

- Baddeley AD, & Hitch GJ (1974). Working Memory In Bower GH (Ed.), *The psychology of Learning and Motivation* (pp. 47–90). New York: Academic Press.
- Bays PM, Catalao RFG, & Husain M (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, 9(10), 1–11.
- Bays PM, & Husain M (2008). Dynamic shifts of limited working memory resources in human vision. *iScience*, i321(5890), 851–854. 10.1126/science.1158023
- Brainard DH (1997). The psychophysics toolbox. *Spatial Vision*, 10, 433–436. [PubMed: 9176952]
- Cousineau D (2007). Confidence intervals in within-subjects designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, 1, 42–45.
- Cowan N (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24, 87–185. [PubMed: 11515286]
- Donkin C, Kary A, Tahir F, & Taylor R (2016). Resources masquerading as slots: Flexible allocation of visual working memory. *Cognitive Psychology*, 85, 30–42. 10.1016/j.cogpsych.2016.01.002 [PubMed: 26794368]
- Dowd EW, Kiyonaga A, Beck JM, & Egner T (2015). Quality and accessibility of visual working memory during cognitive control of attentional guidance: A Bayesian model comparison approach. *Visual Cognition*, 23(3), 337–356. 10.1080/13506285.2014.1003631
- Edin F, Klingberg T, Johansson P, McNab F, Tegnér J, & Compte A (2009). Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences*, 106(16), 6802–6807. 10.1073/pnas.0901894106
- Fougnie D, Cormiea SM, Kanabar A, & Alvarez GA (2016). Strategic trade-offs between quantity and quality in working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 42(8), 1231–1240. 10.1037/xhp0000211 [PubMed: 26950383]
- Klyszejko Z, Rahmati M, & Curtis CE (2014). Attentional priority determines working memory precision. *Vision Research*, 105, 70–76. 10.1016/j.visres.2014.09.002 [PubMed: 25240420]

- Kruschke JK (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, 14(7), 293–300. 10.1016/j.tics.2010.05.001 [PubMed: 20542462]
- Luck SJ, & Vogel EK (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657), 279–281. 10.1038/36846 [PubMed: 9384378]
- Machizawa MG, Goh CCW, & Driver J (2012). Human visual short-term memory precision can be varied at will when the number of retained items is low. *Psychological Science*, 23(6), 554–559. 10.1177/0956797611431988 [PubMed: 22527526]
- Murray AM, Nobre AC, Astle DE, & Stokes MG (2012). Lacking control over the trade-off between quality and quantity in visual short-term memory. *PLoS ONE*, 7(8). 10.1371/journal.pone.0041223
- Oberfeld D, & Franke T (2013). Evaluating the robustness of repeated measures analyses: The case of small sample sizes and nonnormal data. *Behavior Research Methods*, 45(3), 792–812. 10.3758/s13428-012-0281-2 [PubMed: 23184532]
- Palmer J (1990). Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology. Human Perception and Performance*, 16(2), 332–350. [PubMed: 2142203]
- Pelli DG (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10, 437–442. [PubMed: 9176953]
- Rigoux L, Stephan KE, Friston KJ, & Daunizeau J (2014). Bayesian model selection for group studies — Revisited. *NeuroImage*, 84, 971–985. 10.1016/j.neuroimage.2013.08.065 [PubMed: 24018303]
- Roggeman C, Klingberg T, Feenstra HEM, Compte A, & Almeida R (2013). Trade-off between capacity and Precision in Visuospatial Working Memory. *Journal of Cognitive Neuroscience*, 26(2), 211–222. 10.1162/jocn_a_00485 [PubMed: 24047380]
- Spiegelhalter DJ, Best NG, Carlin BP, & Van Der Linde A (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639. 10.1111/1467-9868.00353
- Stephan KE, Penny WD, Daunizeau J, Moran RJ, & Friston KJ (2009). Bayesian Model Selection for Group Studies. *NeuroImage*, 46(4), 1004–1017. 10.1016/j.neuroimage.2009.03.025 [PubMed: 19306932]
- Suchow JW, Brady TF, Fougner D, & Alvarez GA (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, 13(10), 1–8.
- van den Berg R, Awh E, & Ma WJ (2014). Factorial comparison of working memory models. *Psychological Review*, 121(1), 124–149. 10.1037/a0035234 [PubMed: 24490791]
- Vogel EK, McCollough AW, & Machizawa MG (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, 438, 368–387.
- Wilken P, & Ma WJ (2004). A detection theory account of change detection. *Journal of Vision*, 4(12), 1120–1135. 10.1167/4.12.11 [PubMed: 15669916]
- Zhang W, & Luck SJ (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, 453(7192), 233–235. 10.1038/nature06860 [PubMed: 18385672]
- Zhang W, & Luck SJ (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, 453, 233–235. [PubMed: 18385672]
- Zhang W, & Luck SJ (2011). The number and quality of representations in working memory. *Psychological Science*, 22(11), 1434–1441. 10.1177/0956797611417006 [PubMed: 21987693]

Figure 1.

Behavioral task trial sequence. Participants performed a color WM recall task that required remembering 2 or 4 colored squares presented randomly at eight different locations across a 850 ms delay period. Prior to the presentation of the sample display, participants received a cue (2 or 4) indicating the number of memory items to expect on the upcoming trial with 75% accuracy. Participants reported the remembered color of a target item (white square) by selecting the appropriate color on a color wheel presented at test display.

Figure 2.

Results of Experiment 1. *A*: Behavioral estimates of P_m for Cue2 (low-load cue) and Cue4 (high-load cue) at SS2 and SS4. *B*: Behavioral estimates of precision for Cue2 and Cue4 at SS2 and SS4. Error bars represent 95% within-subject confidence intervals (Cousineau, 2007). *C* and *D*: A significant negative correlation between cue-related change in P_m and cue-related change in precision.

Figure 3.

Bayesian model comparison values for each participant and each model. The left side of the figure represents set size 2 conditions and the right side of the figure represents set size 4 conditions. Each column corresponds to one participant and each row corresponds to one of the four models specified. Values depicted in the figure were calculated by assigning a zero value to the best fitting model according to deviance information criterion (DIC) and calculating a relative DIC value for all the remaining models ($DIC_{modelX} - DIC_{modelBest}$). The higher the value of the relative DIC, the worse the fit of the model for that particular subject. Model 0 provided the best fit for a large majority of the participants across both set sizes.

TABLE 1

Model descriptions and Bayesian Model Selection results

Model	Probability of guessing (g) /Capacity (1-g)	Standard deviation (s.d.) /Precision (1/s.d.)	Exceedance probability (P*)		Expectation of the posterior (posterior)	
			SS2	SS4	SS2	SS4
Model 0:						
Constant s.d. & g	Cue4 = Cue2	Cue4 = Cue2	>.999	>.999	.9517	.9505
Model 1:						
Independent s.d. & g	{Cue4}{Cue2}	{Cue4}{Cue2}	<.001	<.001	.0156	.0156
Model 2:						
Ordered s.d. & g	Cue4 < Cue2	Cue4 > Cue2	<.001	<.001	.0163	.0178
Model 3:						
Ordered s.d. & g	Cue4 > Cue2	Cue4 < Cue2	<.001	<.001	.0163	.0161