The Contribution of Attentional Lapses to Individual Differences in Visual Working Memory Capacity

Kirsten C. S. Adam[1*], Irida Mance[1*], Keisuke Fukuda[2], and Edward K. Vogel[1]

Abstract

■ Attentional control and working memory capacity are important cognitive abilities that substantially vary between individuals. Although much is known about how attentional control and working memory capacity relate to each other and to constructs like fluid intelligence, little is known about how trial-by-trial fluctuations in attentional engagement impact trial-by-trial working memory performance. Here, we employ a novel whole-report memory task that allowed us to distinguish between varying levels of attentional engagement in humans performing a working memory task. By characterizing low-performance trials, we can distinguish between models in which working memory performance failures are caused by either (1) complete lapses of attention or (2) variations in attentional control. We found that performance failures increase with set-size and strongly predict working memory capacity. Performance variability was best modeled by an attentional control model of attention, not a lapse model. We examined neural signatures of performance failures by measuring EEG activity while participants performed the whole-report task. The number of items correctly recalled in the memory task was predicted by frontal theta power, with decreased frontal theta power associated with poor performance on the task. In addition, we found that poor performance was not explained by failures of sensory encoding; the P1/N1 response and ocular artifact rates were equivalent for high- and low-performance trials. In all, we propose that attentional lapses alone cannot explain individual differences in working memory performance. Instead, we find that graded fluctuations in attentional control better explain the trial-by-trial differences in working memory that we observe. ■

INTRODUCTION

Individuals with low working memory capacity perform poorly on measures of aptitude such as fluid intelligence and academic achievement (Unsworth, Fukuda, Awh, & Vogel, 2014; Fukuda, Vogel, Mayr, & Awh, 2010; Engle, Tuholski, Laughlin, & Conway, 1999; Turner & Engle, 1989; Daneman & Green, 1986; Daneman & Carpenter, 1980), and extensive work has suggested that individual differences in capacity stem in part from variability in deploying attention (Fukuda & Vogel, 2009, 2011). Low-capacity individuals have specific difficulties in tasks that require attentional control, suggesting that variability in effectively exerting these control mechanisms determines apparent capacity differences. However, because capacity measures are calculated from average performance across an entire session, a common alternative hypothesis remains: Individual differences in capacity are the result of a mixture of "complete attention trials," in which participants maximally allocate their available memory resources, and "lapse trials," in which participants fail to allocate any available memory resources because of complete disengagement from the task. Thus, low-capacity individuals may simply have more lapse trials than high-capacity individuals and consequently show reduced average performance.

A lapse account of working memory performance is consistent with evidence suggesting that task engagement is associated with differences in working memory ability. First, low-capacity individuals engage in mind-wandering more frequently than high-capacity individuals, particularly during cognitively challenging tasks (Mrazek et al., 2012; McVay & Kane, 2010; Smallwood & Schooler, 2006). Second, the slowest RT trials in choice RT tasks are the most predictive of intelligence scores (i.e., the worst performance rule), suggesting that attentional lapses contribute to estimates of individual aptitude (Coyle, 2003). Finally, low-capacity individuals exhibit periods of goal neglect, a performance failure in which task rules are explicitly understood, but are nevertheless not behaviorally executed (Duncan, Schramm, Thompson, & Dumontheil, 2012; Duncan, Emslie, Williams, Johnson, & Freer, 1996). Thus, low-capacity participants may experience an increased number of failures for a variety of reasons, from being captured by distracting internal thoughts to simply giving up on a difficult task.

We have described a lapse model of inattention, in which inattention leads to complete disengagement from the task. However, inattention could also manifest as degraded attentional control rather than as a complete lapse. Under an attentional control model of inattention,

[1]University of Oregon, [2]Vanderbilt University
*These authors contributed equally to this work.

Journal of Cognitive Neuroscience 27:8, pp. 1601–1616
doi:10.1162/jocn_a_00811

impaired attention would lead to reduced performance, though not necessarily to total neglect of the task if the inattention is incomplete. From this view, the efficiency of attentional control may vary over the session in a graded fashion; differences between participants could be viewed as a shift in the distribution of effective attentional engagement. Low-capacity individuals would have a downward shift in this distribution, leading to more frequent periods of partial disengagement than high-capacity participants. Indeed, the evidence reviewed above from mind-wandering frequency and RT variability would be consistent with either complete or graded failures of attention. However, these two models have not been evaluated directly against one another. Critically, most attention and working memory tasks rely on one of two metrics: accuracy and RTs. Accuracy measures are binary (correct or incorrect), making it difficult to test for a graded attentional model. Conversely, RT measures are continuous, making it difficult to test for a complete lapse model. Thus, it is not surprising that the attention literature (predominately RT measures) has strongly assumed a graded model of attention, whereas the working memory capacity literature (predominately signal detection measures) has tested only coarse lapse parameters.

For example, in the change detection paradigm, a randomly chosen item from each memory array is probed, and the participant must indicate whether it is the same as the originally presented item. Here, an individual's capacity is estimated across a series of trials by determining the probability of having stored the probed item on a given trial. Consequently, performance on an individual trial is not informative on its own because an error could be the result of either a complete memory failure (0 items stored) or a successful memory (4 items stored) that was unlucky because the probed item did not happen to be stored. At present, only the complete lapse model has been directly tested in the working memory literature (Van den Berg, Awh, & Ma, 2014; Sims, Jacobs, & Knill, 2012; Morey, 2011; Rouder, Morey, Morey, & Cowan, 2011; Rouder et al., 2008). For example, Rouder et al. (2008, 2011) found that adding a lapse parameter substantially improved model estimates of working memory capacity in a change detection task, particularly by accounting for why errors occur on subcapacity memory arrays. This work demonstrates that an "all or nothing" lapse account could plausibly explain individual differences in capacity. However, because of the aggregate nature of the data, a graded attentional control account could not be evaluated.

Here, we measure how working memory performance fluctuates within a session to determine whether performance failures are better explained by a lapse model (coarse failures) or by an attentional control model (graded failures). To do so, we employ a novel whole-report task that provides a trial-by-trial measurement of the total number of correctly reported items from each array, allowing us to examine the distribution of levels of success on each trial. Our novel whole-report measure is both discrete (each item is correct or incorrect) and continuous (the number of objects correct can fluctuate), allowing us to uniquely distinguish between these two hypothetical models.

Measuring performance fluctuations with continuous whole-report allows us to test distinct predictions made by the competing attentional control and lapse accounts of underperformance. First, although a lapse model predicts that lower-capacity individuals would show higher lapse rates than high-capacity individuals regardless of task load, an attentional control model predicts that substantial performance failures related to individual capacity would only be observed under task loads that necessitate attentional control, such as for supracapacity displays. Second, although a lapse model predicts a bimodal distribution of performance (i.e., lapse vs. full attention), an attentional control model predicts a graded downward shift in performance distributions for low-capacity individuals. After distinguishing between lapse and attentional control models of performance fluctuations, we test several hypotheses about the mechanisms of performance fluctuations. These mechanisms include changes in task engagement over time, task noncompliance, sensory encoding, and attentional control.

In three experiments, we establish the validity of a novel working memory measure and then test hypotheses that differentiate a lapse model and an attentional control model of individual differences in working memory capacity. In a discrete whole-report measure of working memory, participants view a display of brightly colored items, remember the display, and then are asked to identify the color of each item from a fixed set of color choices. Task accuracy is determined by counting the number of correctly identified items. Thus, a participant's level of performance can be calculated for every trial in the experiment. This trial-by-trial measurement of performance is critical for investigating novel hypotheses about the nature of trial-by-trial task engagement.

In Experiment 1a, we tested whether set-size affects the number of performance failures. A coarse lapse model would predict that the amount of time spent disengaged from the task does not vary across set-sizes (Rouder et al., 2008, 2011). Alternatively, an attentional control model would predict an increased rate of performance failures (few items correct) when the set-size exceeds capacity. We had participants complete a change detection memory task and a whole-report memory task for multiple set-sizes (two to six items). We include a change detection measure for two reasons: (1) to initially validate our novel whole-report measure of working memory and (2) as an independent measure that allows us to more closely examine the contributions of within- and between-participant variability in working memory performance while minimizing issues of circularity.

In Experiment 1b, we collected a large number of supracapacity trials to test lapse and attentional control models

and to examine the consistency of performance failures over time. Participants completed a large number of exclusively set-size 6 trials (along with a standard change detection measure). By continuously repeating the same challenging set-size, we could look for fluctuations in engagement over time without confounding carryover effects from conditions with different levels of difficulty.

Finally, we tested whether neural measures of sensory encoding and attentional control predict trial-by-trial working memory performance. In Experiment 2, a new set of participants performed the same tasks as in Experiment 1b while EEG and EOG activity was recorded. Using only the set-size 6 condition was also ideal for EEG analyses, as we could examine fluctuations in neural signals while holding physical stimulation constant. To examine markers of low-performance trials, we analyzed the P1/N1 visual-evoked response and behavioral accuracy for artifact-rejected trials. To examine the potential contribution of attentional control, we measure frontal theta power and posterior alpha power across all time points in the trial.

METHODS

Participants

All participants gave written informed consent according to procedures approved by the University of Oregon institutional review board. Participants were compensated for participation with course credit or monetary payment ($8/hr for behavior, $10/hr for EEG). A unique set of individuals participated in each experiment, with 40 in Experiment 1a, 45 in Experiment 1b, and 26 in Experiment 2. One participant from Experiment 1b was excluded from analyses for failure to comply with task instructions. After artifact rejection, three participants were excluded from Experiment 2 analyses for artifact rates in excess of 25% or fewer than 300 remaining trials, and one participant did not complete the change detection measure, so they were excluded from between-participant analyses using change detection but included in within-participant analyses.

Stimuli

Stimuli were generated in MATLAB (The MathWorks, Natick, MA) using the Psychophysics toolbox (Brainard, 1997; Pelli, 1997) and presented on 21-in. CRT monitors. Participants were seated approximated 60 cm from the monitor. In Experiment 1a, stimuli (2.50° visual angle) for both whole-report and change detection tasks consisted of eight colors (RGB values: Red = 255 0 0; Green = 0 255 0; Blue = 0 0 255; Magenta = 255 0 255; Yellow = 255 255 0; Cyan = 0 255 255; White = 255 255 255; Black = 0 0 0), presented on a gray background (RGB = 128 128 128). In Experiment 1b and Experiment 2, one additional color (RGB: Orange = 255 128 0) was added to the potential memory set colors.

Tasks

Change Detection Task

The change detection task used in all experiments followed standard procedures (Luck & Vogel, 1997). In Experiment 1a, participants were presented with arrays of two to six colored squares for 150 msec (memory array), which disappeared for 900 msec (retention period), followed by the presentation of one colored square (test probe) at the location previously occupied by an item in the memorandum. In Experiment 1b and Experiment 2, participants were presented with arrays of four, six, or eight items, and trials used a 250-msec stimulus array and 1000-msec retention period. On 50% of trials, the test item was the same as the item presented in the memory array, and in the remaining 50% of trials, the test item was different. Participants were instructed to make an unspeeded button press to indicate whether the color of the probe had changed. The next trial began after an intertrial interval of either 900 msec (Experiment 1a) or 1000 msec (Experiment 1b and Experiment 2).

Whole-report Task

The whole-report procedure was similar to the change detection task with the exception that individuals recalled each item shown in the memory array. At response, individuals were shown a three by three matrix of colors over each location of memory array items; they were instructed to use a mouse to click on the color square corresponding to the memory array item at each location. In Experiment 1a, individuals were encouraged to respond to as many items as they could remember and advanced to the next trial by pressing the spacebar. In Experiments 1b and 2, individuals were required to respond to all items in the memory array. The next trial began when all responses were made (Experiment 1b) or after the participant clicked to indicate they were ready for the next trial (Experiment 2). Stimulus timing parameters were the same as for the respective change detection task, except for Experiment 2. In Experiment 2, the retention interval and intertrial interval periods were increased to 1300 msec to provide a larger time window for oscillatory analyses.

Procedures

Experiment 1a

Participants completed two blocks of 150 trials of the change detection task and the whole-report task across several set-sizes (two to six items). Set-sizes were intermixed within blocks for all experiments. For each of

the two tasks, participants completed 150 trials, for a total of 30 trials per set-size.

Experiment 1b

Participants performed five blocks of 36 trials of the change detection task, for a total of 60 trials per set-size. For the whole-report task, participants completed 10 blocks of 30 trials (300 trials total), and all arrays were set-size 6.

Experiment 2

Participants performed the same tasks as in Experiment 1b, while we recorded EEG activity. Participants performed five blocks of 36 trials of the change detection task, for a total of 60 trials per set-size. For the whole-report task, participants completed 15–18 blocks of 30 trials (450–540 trials total), and all arrays were set-size 6.

EEG Data Collection

EEG activity was recorded at 20 electrodes mounted in an elastic cap (ElectroCap International, Eaton, OH) using our standard recording and analysis procedures (McCollough, Machizawa, & Vogel, 2007). The International 10/20 sites F3, FZ, F4, T3, C3, CZ, C4, T4, P3, PZ P4, T5, T6, O1, and O2 were used along with five nonstandard sites: OL midway between T5 and O1, OR midway between T6 and O2, PO3 midway between P3 and OL, PO4 midway between P4 and OR, and POz midway between PO3 and PO4. All sites were recoded with a left-mastoid reference, and the data were re-referenced offline to the algebraic average of the left and right mastoids. Horizontal EOG was recorded from electrodes placed ~1 cm to the left and right of the external canthi of each eye to measure horizontal eye movements. To detect blinks, vertical EOG was recorded from an electrode mounted beneath the left eye and referenced to the left mastoid. The EEG and EOG signals were amplified with an SA Instrumentation amplifier (Fife, Scotland) with a bandpass of 0.01–80 Hz and were digitized at 250 Hz in Labview 6.1 running on a PC.

Trials including blocking, blinks, or large (>1°) ocular movements were rejected and excluded from further analysis. For ERP analyses, we baselined the signal over the 200 msec prior to the timelocking event (onset of the memory array); trials were filtered with a low-pass finite impulse response filter with a cutoff of 40 Hz. For oscillatory analyses, we bandpass-filtered the raw EEG using a two-way, least-squares finite impulse response filter using the eegfilt.m filter function from the EEGLAB Toolbox (Delorme & Makeig, 2004; Brainard, 1997). We applied the MATLAB Hilbert transform (hilbert.m) to extract the instantaneous power values. For spectrograms, power data were calculated separately for each 1 Hz band. For band-specific analyses, power data were calculated for typically defined frequency bands (theta: 4–7 Hz; alpha: 8–12 Hz; beta: 13–22 Hz).

Statistical Analyses

Change Detection Capacity

Each participant's change detection accuracy was transformed into a K estimate using Cowan's formula $K = N \times (H - FA)$, where N represents the set-size, H is the hit rate, and FA is the false alarm rate (Cowan, 2001). In Experiment 1a, the average of set-sizes 4, 5, and 6 arrays were used to estimate each participant's change detection capacity. In Experiment 1b and Experiment 2, all set-sizes 4, 6, and 8 were used to estimate capacity.

Whole-report Accuracy

For the whole-report procedure, we estimated accuracy in two ways. First, we calculated the average number of correctly reported items per set-size. Second, we split performance into the proportion of trials in which participants correctly reported zero, one, two, and so forth, for each set-size. This method allowed us to measure the proportion of trials during which participants exhibited impaired working memory performance and how failures varied as a factor of set-size and working memory capacity. By examining trial-by-trial accuracy, we can observe the impact of performance fluctuations on overall working memory ability.

Simulation Analyses

The greater number of trials in Experiment 1b allowed us to perform a finer analysis of trial-by-trial performance. First, we characterized the expected performance outcome if participants had a complete attentional lapse. We ran 30 iterations of a simulation where the computer guessed colors (without replacement) for the six items in the display across 300 trials. We also wanted to characterize the effect of guessing inflation on performance, especially for high-performance trials. To do so, we assigned three correct objects to each trial in the simulation and examined the effect of guessing colors without replacement for the remaining three items. Finally, we ran simulations to test whether a complete lapse model or a graded attentional control model could explain trial-by-trial fluctuations in working memory performance. The two hypothetical models both predict that trial-by-trial performance in the whole-report task can be modeled by (1) an upper limit on total available working memory resources and (2) a probability of allocating the available working memory resources. The lapse and attentional control models differ only in the higher-order distribution used to predict the probability of allocating working memory resources on a trial-by-trial basis. Importantly, the parameter values for the models were chosen only

with respect to mean performance; the model-fitting procedure was blind to the underlying distribution of trials. After the model that best fit the mean was chosen (minimum difference between true and simulated mean), we used the residuals of the model fits (root mean square error [RMSE]) to test the fit of each lapse model across participants.

Lapse Model

The lapse model was based on the assumption that individuals are either fully on task or completely disengaged from the task (Van den Berg et al., 2014; Morey, 2011; Rouder et al., 2008, 2011). This model predicts that during full engagement trials, participants will be able to maximally allocate working memory resources (maximum capacity), with guessing inflation accounting for the trials in which the number of items correct exceeds this capacity. For fully disengaged trials, the model predicts that responses will only be based on guessing. The higher-order distribution of the complete lapse model is Bernoulli-distributed, a distribution with only two values, zero and one. The proportion of zeroes in the distribution represents disengaged trials, and the proportion of ones represents fully engaged trials. On each simulated trial, one value is pulled from the higher-order distribution and multiplied by the maximum capacity parameter to yield the performance outcome. For example, if maximum capacity parameter is three items, then on a trial where a "1" is pulled, the participant achieves 1×3, or three items. Guessing is accounted for the remaining items in the set-size (in this case, three guesses). For each participant and parameter level, we simulated 300 trials for each participant. During each run of the simulation, maximum capacity was initially held constant at three items, whereas the proportion of lapses (zeroes) in the higher-order distribution was parametrically increased from 0% to 100% in steps of 1%. We allowed for guessing by randomly drawing without replacement from the nine possible colors. For example, for a lapse parameter of 20%, we sampled only from the random guessing distribution for 20% of the trials, whereas for the remaining 80% of trials, capacity was set to three items plus three guesses. The simulated mean number correct for each run was compared to each individual's empirical mean number correct to determine the best-fit lapse parameter. This was repeated for each participant in the data set. The lapse parameter that best fit the mean experimental performance for each participant was used to generate the underlying response distributions for the model. The aggregate of the generated response distributions were then used to test the fit of the complete lapse model.

Attentional Control Model

The attentional control model was similar to the complete lapse procedure with the exception that the higher-order distribution was beta-distributed. Like a Bernoulli distribution, a beta distribution has values bound between zero and one. However, the beta distribution contains many graded outcomes between zero and one. In the attentional control model, one represents maximal attentional engagement and zero represents minimal attentional disengagement. We used a beta distribution because we could model the graded probability of maximally allocating working memory resources on each trial. The α parameter of the beta distribution was parametrically varied from zero to six in steps of .01, whereas the β parameter was constrained to one. Thus, only a single parameter was varied to shift the relative proportion of more engaged and less engaged trials. For each cycle of α values, we randomly generated a value from a beta distribution of $[\alpha,1]$ for each trial. On each trial of the simulation, one value was pulled from the beta distribution (e.g., 0.5), multiplied by the maximum capacity parameter (e.g., $0.5 \times 3 = 1.5$), and rounded to the nearest discrete outcome (e.g., 2). As in the complete lapse model, for each participant in the data set, the α value that best-fit the participant's observed data was selected. These best-fit α values were then used to generate a distribution of responses predicted by the model, which were then used to test the fit of the model by calculating the RMSE.

RESULTS

Mean Whole-report Performance Corresponds with Change Detection Capacity

The task and results from Experiment 1a are shown in Figure 1. For both tasks, working memory performance increases with set-size, reaching a stable plateau around three to four items. Averaged across set-sizes, the mean change detection capacity estimate (Cowan's K) was 2.62 ($SD = 0.72$), and the mean number of items correct on the whole-report task was 2.91 ($SD = 0.51$). We ran separate repeated-measures ANOVAs to verify the change in performance across set-sizes for change detection and for whole-report performance. Change detection performance was significantly different across set-sizes, $F(2.48, 96.68) = 18.13$, $p < .01$. Here and for other cases in which Mauchly's test indicated a violation of the assumption of sphericity, we report Greenhouse–Geisser corrected F statistics and p values. We also ran planned simple contrasts (comparing all smaller set-sizes to the largest set-size) to check for a plateau in performance; we found that the only significant contrast was between set-size 2 and set-size 6, $F(1, 39) = 18.13$, $p < .01$. The comparisons between set-size 6 and the other set-sizes (3, 4, and 5) were not significant, all comparisons $p > .20$, suggesting that performance reached a plateau at set-size 3. Whole-report performance was also significant across different set-sizes, $F(1, 39) = 89.54$, $p < .01$. We again ran planned simple contrasts comparing all lower

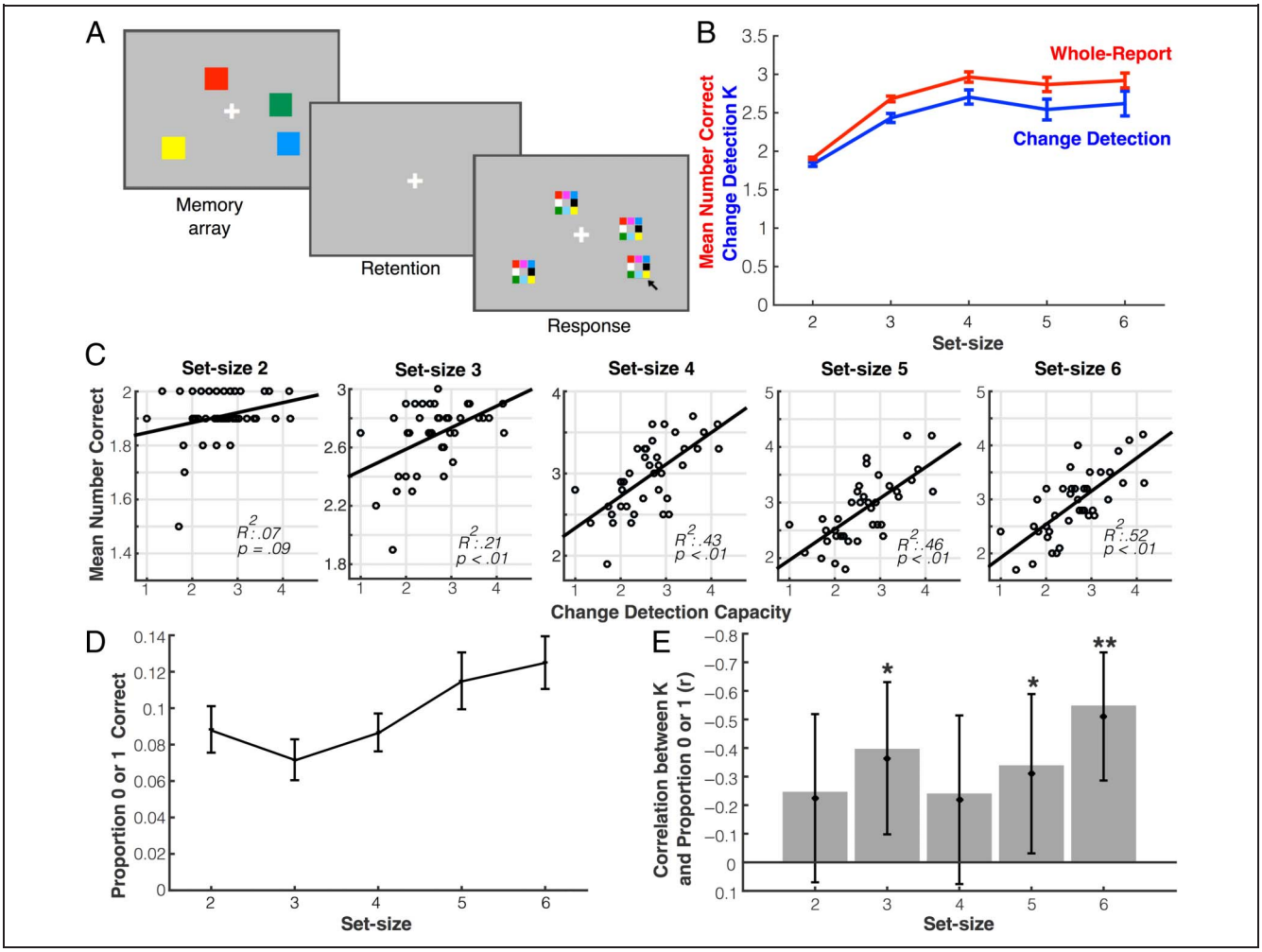

Figure 1. Results from Experiment 1a. (A) Illustration of the task design and stimuli in Experiment 1a. (B) Overall performance changes in a similar manner for change detection (blue) and whole report (red) across set-sizes. (C) The correlation between mean whole-report performance and change detection capacity at each set-size. The proportion of performance failures in whole-report (0 or 1 correct) increases across set-size (D) and explains more variance in capacity across set-size (E).

set-sizes to set-size 6. We found that set-size 6 performance was significantly higher than set-size 2, $F(1, 39) = 119.49$, $p < .01$, and set-size 3, $F(1, 39) = 9.34$, $p < .01$, but not for other set-sizes ($p > .30$), suggesting that performance reached a plateau around set-size 4.

In addition to examining how within-task performance changes with load, we can also examine if the relationship between whole-report performance and change detection K is consistent across loads. For each whole-report set-size, we calculated the correlation with a typical measure of change detection capacity (set-sizes 4, 5, and 6) and the mean number of items correctly reported for each set-size. The relationship between change detection K and whole-report performance across set-sizes is shown in Figure 1B. At the lowest set-size (2 items) the relationship between the two measures was nonsignificant; the relationship becomes significant at set-size 3 and increases in strength as participants become more overloaded with items (Figure 1C). However, the lack of correlation between change detection capacity and whole-report performance at set-size 2 could be due to ceiling effects; most participants were nearly perfect at responding to the set-size 2 trials. In summary, whole-report and change detection estimates of average working memory performance are strongly related to each other. Next, we can investigate whether specific performance outcomes (e.g., 0 items correct) change across working memory loads and whether performance outcomes also predict individual differences in working memory capacity.

The Relationship between Performance Failures and K Is Set-size Dependent

To assess performance failures, we measured the proportions of trials in which a given number of items were correctly identified on each trial. We defined performance failures as trials in which participants scored 0 or 1 items correct out of 6 possible items, because a simulation of guessing yielded 0 or 1 correct on 85% of trials

(see simulation results; Figure 3A). In particular, we were interested in quantifying whether the proportion of extreme performance failures was constant or variable across set-sizes. As illustrated in Figure 1C, we found that the proportion of performance failures increased across set-sizes. Previous lapse models have assumed that the rate of performance failures is constant across set-size. Instead, a repeated-measures ANOVA reveals that there is a significant difference in performance failures between set-sizes, $F(3.33, 130.02) = 5.16$, $p = .001$, with performance failures increasing across set-size (Figure 1D). We ran planned simple contrasts to test how the rate of performance failures at earlier set-sizes compared to the rate at the highest set-size (6 items). We found that the only nonsignificant contrast was for the rates at set-size 5 and set-size 6, $F(1, 39) = 0.44$, $p = .51$. All other set-sizes had failure rates lower than set-size 6, minimum difference $p < .03$. Additionally, the relationship between change detection capacity and set-size 6 performance failures is the strongest, $r = -.55$, $p < .01$, 95% CI [−0.73, −0.29], (Figure 1E).

This set of correlations reveals that, although performance failures occur at all set-sizes, they are consistently diagnostic of individual differences in capacity only for supracapacity set-sizes. Thus, although all participants perform very poorly on a subset of trials, low-capacity individuals display much greater proportions of poor performance trials. Furthermore, this difference between high- and low-capacity participants emerges only for supracapacity arrays, supporting an attentional control model over a lapse model. Given these findings, we next examined performance for a task where participants repeatedly performed set-size 6 trials; this allowed for a more precise characterization in performance distributions and how performance may change over time.

Fluctuations in Whole-report Performance Predict Change Detection Capacity

In Experiment 1b, a new sample of participants completed 300 trials of set-size 6, allowing us to examine fluctuations in task performance that are independent of trial-by-trial variability in task difficulty. The mean change detection K was 2.90 ($SD = 0.98$), and the mean whole-report accuracy was 2.87 ($SD = 0.49$). Again, we found a strong positive relationship between change detection K and overall whole-report performance, $r = .55$, $p < .01$ 95% CI [0.30, 0.73] (Figure 2A).

We initially examined individual differences in a coarse manner by splitting participants into three groups based on their change detection performance (Figure 2B). Participants in the low-K group had change detection scores more than one standard deviation below the mean K score. Likewise participants in the high-K group had change detection scores more than one standard deviation above the mean K score. All other participants were placed in the middle K group. As can be seen by the distributions, the prevalence of performance failures (0, 1, or 2 correct) increases across these performance groups. A simple lapse model of performance would predict bimodality, with large proportions of trials at 0 and at typical capacity values. Here, the low-capacity group has more complete failures (0 or 1 correct) than the high-capacity group, but neither group shows bimodality. Instead, this difference in performance failures appears to be part of an overall shift in performance distributions that is more consistent with an attentional control model of individual differences. We observe a downward shift in performance distributions for low-K participants and an upward shift for high-K participants. However, although using an extreme groups split is useful for summarizing gross differences between groups, such an approach often creates statistical problems (Conway et al., 2005). As such, we also wanted to examine the fine-grained, correlational differences between participants' distributions.

To better visualize how subtle distributional shifts correspond with visual working memory capacity, we can plot distributions from all participants in a single heat map and sort the rows by change detection K (Figure 2C). The heat map depicts the distribution of number correct for all participants. Each horizontal line represents a different participant; the lines are arranged along the y axis according to each participant's change detection score. The x axis represents different trial outcomes, between 0 and 6 correct. Intensity of the heat scale represents proportion of trials that fall into each score category. Here, we can see a strong, dark band at three items, indicating that most participants had a larger proportion of trials in which they scored 3 correct. Again, none of the participants show a pattern that is consistent with a bimodal lapse model of performance. To quantify the relationship between K and performance outcomes, we present the correlation values for each level of performance (Figure 2D). As suggested by the consistent gray band at 3 items correct, the correlation between number of three correct trials and K was nonsignificant ($r = .02$, $p = .89$, 95% CI [−0.27, 0.31]). On the other hand, the number of correct objects in categories above ($r = .54$, $p < .001$, 95% CI [0.66, 0.88]) and below ($r = -.52$, $p < .001$, 95% CI [−0.71, −0.26]) this mode strongly predicted K.

Monte Carlo Simulation of an Attentional Control Model Is Better than a Lapse Model

The greater number of trials in Experiment 1b allowed us to use simulation approaches to test potential models of individual differences in performance. In particular, we were interested in whether performance failures are better characterized as an all or nothing engagement in the task (lapses) or as varying degrees of engagement in the task (attentional control). For the purposes of modeling, the size of our maximum resource pool is described in items, but this is only because of the nature

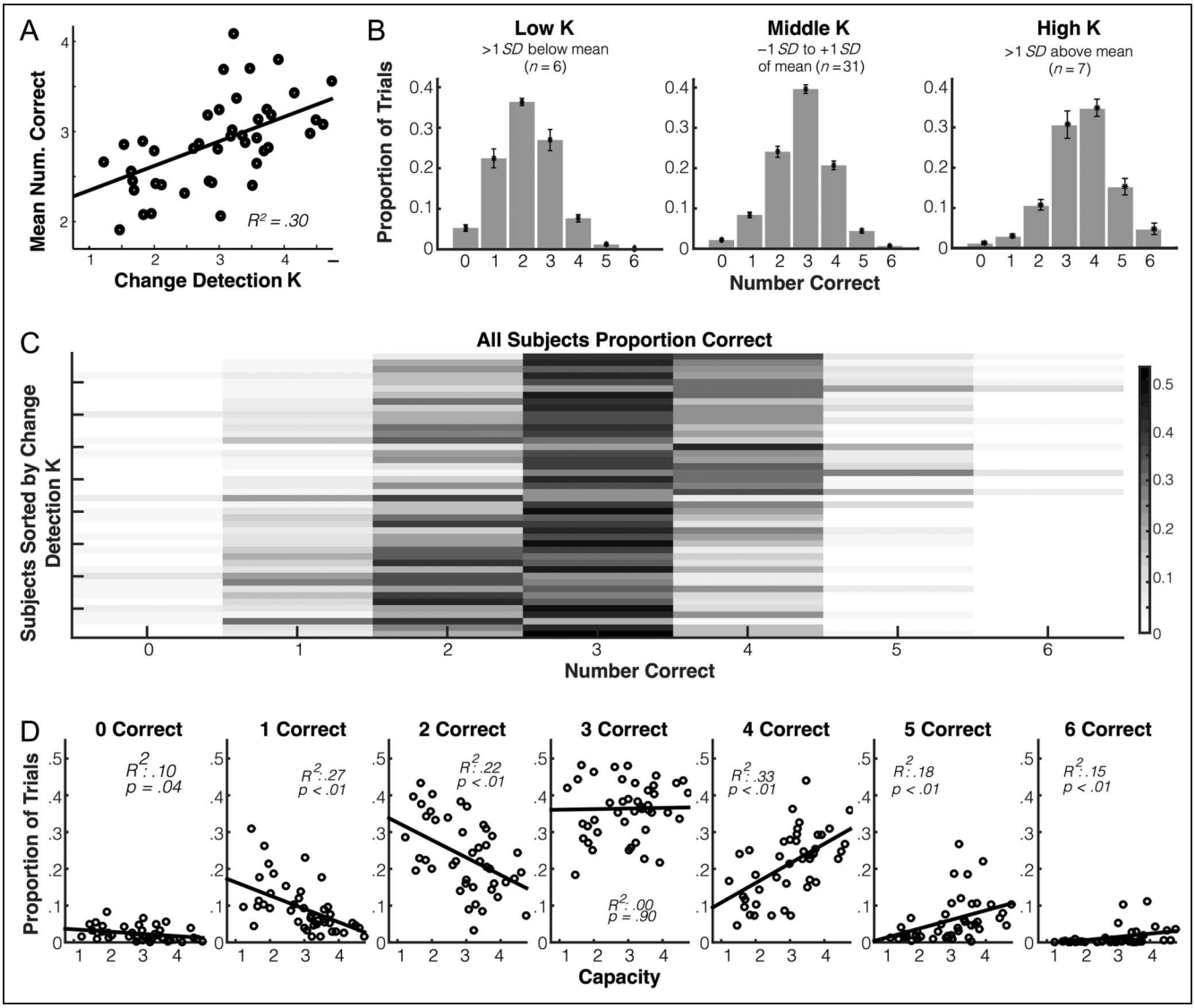

Figure 2. Whole-report performance distributions for Experiment 1b. (A) Correlation between mean whole-report performance and change detection capacity in Experiment 1b. (B) Performance distributions for participants split into extreme groups by their change detection score. (C) Performance distributions for all participants in Experiment 1b. Each column represents the performance outcome (number of items correct for a given trial), and each row represents a participant (sorted by capacity). Differences between participants are best characterized as a subtle upward or downward shift of the performance distribution (with a central tendency at 3 for most participants). (D) Performance outcomes correlate with change detection estimates of capacity for all levels of performance except for 3 correct.

of our behavioral assay. Our task necessarily involves a discrete outcome (a discrete number of items correct out of six), and all of our modeling efforts rely upon such estimates of trial-by-trial performance. However, this experiment cannot speak to any of the current debates about the nature of the limit on working memory. We make no claims that discrete slots models are preferred over continuous resources models on the basis of these data. Instead, we are interested in tracking variations in a participant's typical performance level (the deployment of their resources, whatever the underlying structure of the resources). Indeed, it has been proposed that both discrete and continuous models of working memory resources could plausibly implement a resource that varies from trial to trial (Van den Berg et al., 2014).

Lapse Performance and Guessing Inflation

Before testing lapse and attentional control models, we first characterized the performance outcomes for guessing among six items and the effects of guessing inflation. Guessing without replacement for set-size 6 yielded 0 or 1 correct 85% of the time for nine possible colors (Figure 3A). This means that, on the remaining 15–17% of guessing trials, a participant may have reported two or more items correct, even when they truly had zero items in mind. Given our simulation results, we used 0 or 1 correct as a conservative definition of performance failures for all analyses. We also ran a simulation to account for the effects of guessing inflation given knowledge about three items and guessing without replacement. The

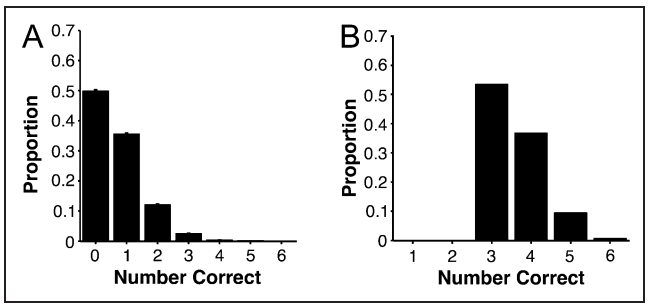

Figure 3. Monte Carlo simulation of lapse performance and guessing inflation. (A) Results from a simulation of guessing without replacement from nine colors over six objects. (B) Results from a simulation of guessing inflation when participants get 3 items correct and guess without replacement from the remaining colors over 3 objects.

guessing inflation distribution has a strong peak at 3 but also has a large number of trials where participants get 4 or 5 items correct by chance (45%; Figure 3B). Thus, guessing inflation can account for a large percentage of above 3 trials for a participant with a true maximum capacity of 3, and it is important to control for this effect in any simulation model.

Testing Lapse and Attentional Control Models

The lapse model specified that lapse events occur as a total loss of attentional engagement, whereas the attentional control model specified that lapse events occur as a variable loss of attentional engagement. Using a maximum capacity parameter of 3 items correct, our results show that we could successfully recreate the observed mean whole report performance with both the lapse model (R^2 = .95, p < .01, 95% CI [0.90, 0.97]) and the graded attentional control model (R^2 = .82, p < .01, 95% CI [0.70, 0.90]. However, only the attentional control model reliably fit the observed distribution of responses (mean RMSE = 0.14 [0.01]; values in brackets represent *SEM*). The failure of the complete lapse model (mean RMSE = 0.29 [0.01]) was due to an overestimation of the proportion of trials in which individuals reported zero or three and an underestimation of the proportion of trials in which individuals reported two items, thus producing a bimodal distribution of expected responses. This difference in model fit was significantly different, $t(43) = -8.5$, $p = 4.5 \times 10^{-11}$, 95% CI [−0.11, −0.17]. Additionally, we found that the attentional control model is better than the complete lapse model for low-K participants, $t(5) = -14.1$, $p = 1.64 \times 10^{-5}$, 95% CI [−0.33, −0.24], and middle-K participants, $t(30) = -8.34$, $p = 1.3 \times 10^{-9}$, 95% CI [−0.17, −0.10], although neither model was good for extremely high-K participants, $t(6) = -0.51$, $p = .31$, 95% CI [−0.07, 0.03] (Figure 4).

Next, we wanted to test whether changing the maximum capacity parameter in the attentional control model would result in increased fit for a different participant. In particular, we noticed that the high-K group was fit poorly

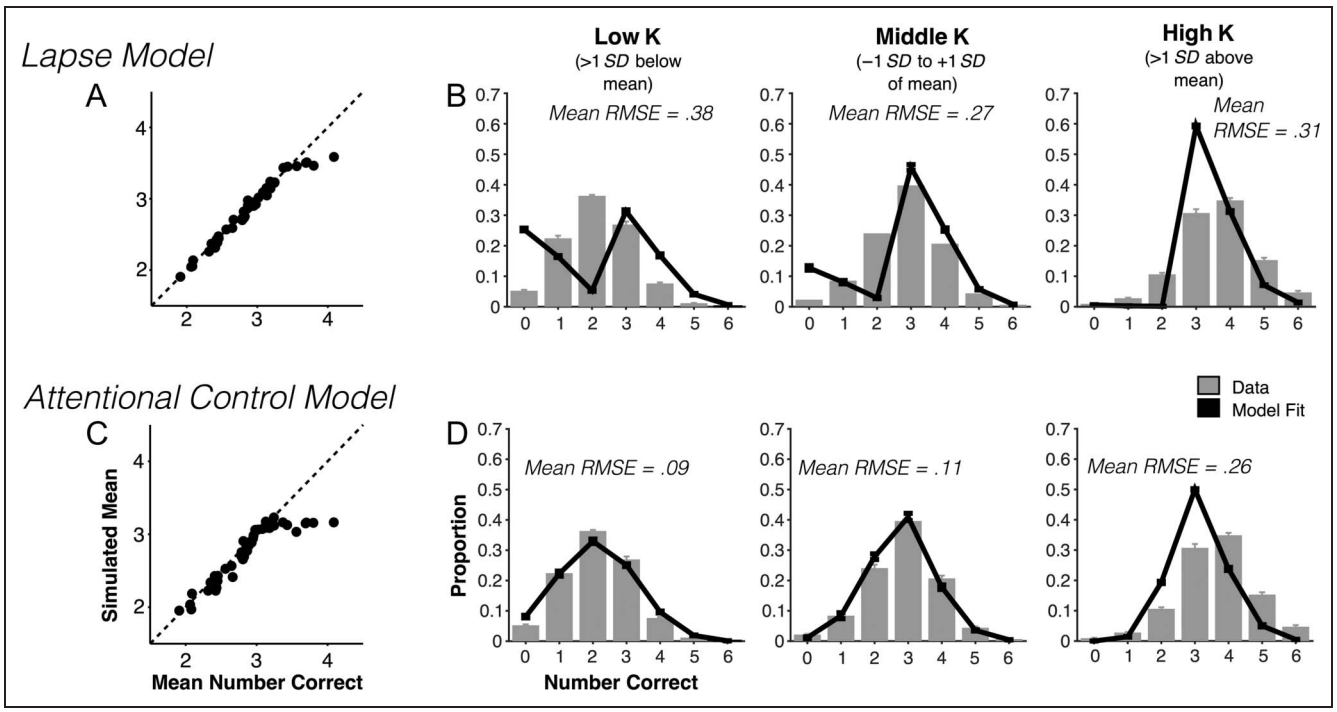

Figure 4. Monte Carlo simulation results for lapse and attentional control models of performance fluctuations. (A) The simulated mean number correct from the lapse model as a function of the actual mean number correct. (B) Data (gray bars) and lapse model fits (black lines) from the extreme groups split of participants. (C) The simulated mean number correct from the attentional control model as a function of the actual mean number correct. (D) Data (gray bars) and attentional control model fits (black lines) shown over the extreme groups split of participants.

by a maximum capacity of three items (for either model). Additionally, we thought that participants in the low-K group, with an overall downward shifted distribution, may be better fit by a maximum capacity of two items. We found that changing the maximum capacity parameter from three to two items did not improve fits for the low-K group (ΔRMSE = +0.04 [0.04], p = .30) and significantly decreased fits for the other two groups (ΔRMSE = +0.21 [0.02], $p = 6 \times 10^{-10}$, and +0.23 [0.03], p = .001, respectively). Next, we tested if increasing the maximum capacity parameter would increase fits for the high-K group. This model reproduced means for participants, R^2 = 0.96, $p < .01$, 95% CI [0.92, 0.98]. Increasing the maximum capacity parameter from three to four improves fits for the high-K group (ΔRMSE = −0.19 [0.04], p = .005) but significantly decreases fits for the low-K (ΔRMSE = +0.09 [0.02], p = .005) and middle-K (ΔRMSE = +0.03 [0.01], p = .04) groups. Finally, we tested a model in which there is no capacity maximum (maximum capacity parameter is six items). We found that this model accurately corresponds to mean performance, R^2 = 0.91, $p < .01$, 95% CI [0.83, 0.95], but resulted in poor fits that were significantly worse than the proposed limited capacity model (ΔRMSE = +0.16 [0.02], $t(43) = -8.66$, $p = 3 \times 10^{-11}$, 95% CI [−0.19, −0.12]).

Performance Fluctuations Occur Consistently over Time

One alternative explanation for the increased prevalence of performance failures for low-K participants is that they took much longer to learn the task and had an inflated level of performance failures in early blocks. Similarly, the relationship between performance failures and capacity could also be explained if low-K participants "give up" at the end of the experiment. To test these time-based explanations of performance, we examined the occurrence of performance failures over time.

In Figure 5A, we illustrate performance for all participants and all trials across time. Each row represents a participant, and the rows are sorted by overall whole report performance. As such, low-performing participants are on the bottom of the graph, and high-performing participants are on the top. Black tick marks represent extreme performance failures (0 or 1 correct), gray tick marks represent below a typical modal performance (2 correct), and white tick marks represent full engagement trials (3 or more correct). Red vertical lines represent block breaks. Because there was only condition in Experiment 2 (set-size 6), there is a unique opportunity to look at variations in working memory performance, on a single

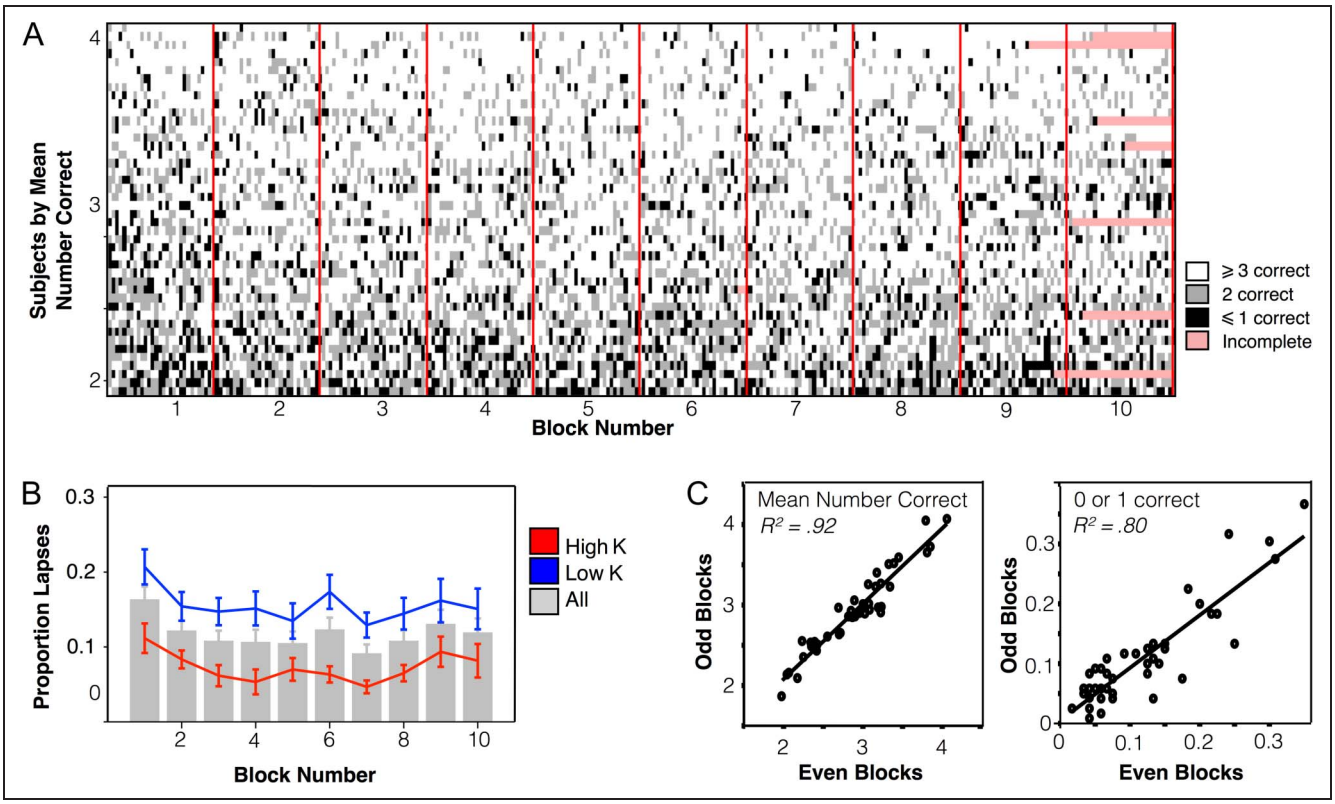

Figure 5. Performance fluctuations over time in Experiment 1b. (A) Data from all participants and all trials is shown over time. Each subject is a row (sorted by whole-report performance), and the horizontal axis corresponds to trials over time. Tick mark color corresponds to the performance outcome; red lines delineate block breaks. (B) A summary of the prevalence of performance failures (0 or 1 items correct) over blocks. Gray bars show the mean performance level; red and blue lines illustrate a median split of subjects. (C) Reliability of mean number correct (left) and prevalence of performance failures (right) for even versus odd blocks.

trial basis, that are not due to condition or set-size difficulty. Instead, fluctuations in performance represent fluctuations in task engagement. As can be seen by the heterogeneous appearance of black and white tick marks, performance failures are scattered throughout the experiment for both high and low-K participants.

To quantify this trend, we examined the frequency of performance failures over time (Figure 5B). We performed a median split based on change detection K and ran a two-way repeated-measures ANOVA with Run (9 blocks, within participants) and Group (2 groups, between participants) as the main factors. We plotted average accuracy for all participants and blocks, but we restricted the ANOVA to Blocks 1–9 because some participants did not finish Block 10 (those with pink tick marks in Figure 5A). We found a significant main effect of Group, $F(1, 42) = 132.5$, $p < .001$, and of Block, $F(8, 336) = 4.96$, $p \leq .001$ but no interaction of Block and Group, $F(8, 336) = 0.677$, $p = .71$. The lack of interaction indicates that, although there were block effects for both groups of participants, the difference between high- and low-capacity participants was consistent over time. Post hoc pairwise comparisons revealed that only the first block was significantly different from any of the other blocks. Lapses were significantly higher in the first block than in Blocks 2 through 5 ($p < .015$) or in Blocks 7 through 8 ($p < .02$). Notably, participants did not complete a set of practice trials before beginning the experiment, so a learning effect would be expected. In summary, an increased lapse rate in the first block explains the difference in lapses over time, but a differential ability to learn the task does not explain the consistent difference in lapses between participants.

As an additional check on the reliability of whole-report performance within a session, we performed a between-block reliability analysis. The reliability of average performance and performance failures is shown in Figure 5C as the correlation between performance on even blocks and odd blocks. Because not all participants finished 10 blocks, we restricted the analysis to Blocks 2–9 and included all participants. We used Cronbach's alpha to quantify reliability. Performance failures were highly reliable for both mean performance (Cronbach's alpha = .97, for Blocks 2–9) and for performance failures (Cronbach's alpha = .93, for Blocks 2–9). In summary, we show that differences in the preponderance of performance failures are not due to learning differences between high- and low-capacity participants. Furthermore, whole-report estimates of working memory performance are highly reliable throughout the session.

Performance Fluctuations Are Not Due to Artifacts or Sensory Encoding Differences

In Experiment 2, we recorded EOG and EEG while participants completed the whole-report task. First, we wanted to examine the role of simple task noncompliance

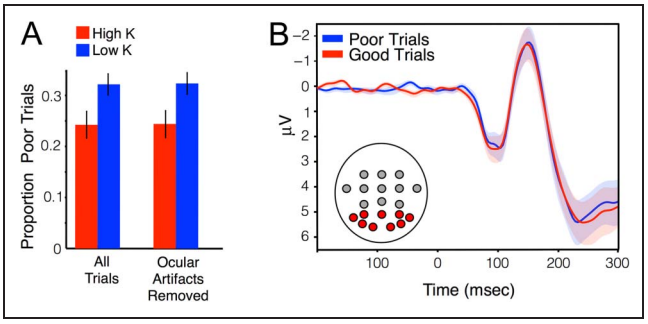

Figure 6. Performance fluctuations do not relate to task compliance or sensory encoding. (A) The prevalence of poor trials (2 or fewer correct) before and after removing ocular artifacts. (B) The P1/N1 visual-evoked response as a function of performance outcome (good versus poor trials).

on the rate of poor performance trials (less than 3 items correct). We instructed participants to keep their eyes on a fixation cross and not to close their eyes during the presentation of the memory array; if participants did not follow these instructions (e.g., moving their eyes away from the screen, blinking during the presentation array), then they may show degraded performance. To test this possibility, we measured the occurrence of poor performance trials before and after excluding trials containing ocular artifacts. We found that the overall ratio of low-performance trials was not significantly changed after removing ocular artifacts, $t(22) = 1.2$, $p = .24$ (Figure 6A), and the relationship between poor performance trials and change detection K was preserved, $R^2 = .23$, $p = .022$. Furthermore, the relationship between the percentage of artifact-rejected trials and K was nonsignificant, $R^2 < .01$, $p = .85$. Thus, for the vast majority of lapse trials, the participant's eyes are indeed open and pointed toward the screen. The percentage of trials that participants are negligent of eye movement instructions does not predict their overall performance level.

Next, we wanted to examine whether or not poor performance trials were associated with decreased sensory processing (Weissman, Roberts, Visscher, & Woldorff, 2006). We measured the mean amplitudes of the visual-evoked P1/N1 ERP components and found that there were no significant differences in amplitude between poor trials and good trials (greater than 3 items correct) for the P1 (70–120 msec), $t(22) = 0.59$, $p = .56$, or the N1 (130–170 msec), $t(22) = 0.49$, $p = .63$ (Figure 6B). Together, we find no evidence that participants show decreased early sensory processing of external stimuli during poor performance trials.

Performance Fluctuations Are Related to Frontal Theta and Posterior Alpha Power

Finally, we tested whether hypothesized neural correlates of attentional engagement and working memory predicted whole-report performance. In particular, we focused on

spectral power in the theta and alpha frequency bands. Frontal theta power has been shown to relate to measures of executive control (Cavanagh & Frank, 2014), working memory load (Deiber et al., 2007; Jensen & Tesche, 2002), successful retrieval (Hsieh & Ranganath, 2014), and manipulation of information in working memory (Itthipuripat, Wessel, & Aron, 2013). Decreased alpha power has been shown to relate to attention and semantic memory performance (Klimesch, 1999; Klimesch, Doppelmayr, Schimke, & Ripper, 1997) and with task difficulty and working memory load (Stipacek, Grabner, Neuper, Fink, & Neubauer, 2003; Gevins, 2000; Gevins, Smith, McEvoy, & Yu, 1997). Here, we wanted to examine whether, given the same difficult task load, we could predict trial-to-trial fluctuations in participants' success using markers shown to be related to overall task difficulty.

First, we wanted to test the simple hypothesis that average frontal theta or posterior alpha power (calculated over the entire trial period) correlates with overall working memory ability. Although some studies have found evidence for a relationship between individual differences in spectral power and cognitive ability (Zakrzewska & Brzezicka, 2014; Gevins, 2000), such relationships are more often unreported in the literature. In our study, we found no relationship of mean posterior alpha power with either change detection ($r = .07$, $p = .88$) or whole-report performance ($r = -.24$, $p = .26$). We found that frontal theta power correlates positively with change detection capacity ($r = .45$, $p = .03$), consistent with previous measures of theta power and working memory. However, theta power did not correlate with whole-report performance ($r = .28$, $p = .19$), despite the strong relationship between change detection and working memory performance for this sample ($r = .61$, $p < .01$). In summary, we conclude that overall frontal theta power may relate to working memory performance in some settings, but either (1) it does not consistently relate to all working memory tasks or (2) the power in this sample is insufficient to show the relationship consistently across tasks.

Despite inconsistent findings for a strong between-participant relationship of working memory ability and theta, we are in a good position to examine within-participant, trial-by-trial predictors of performance. In the EEG experiment, participants completed many trials of the same condition; as such, we have a large number of trials to examine oscillatory predictors of trial-by-trial performance. In Figure 7, each subplot represents the time–frequency plot for a single electrode during Experiment 2. The heat map is created by subtracting the time–frequency plot for poor trials (≤2 correct) from good trials (≥4 correct). A difference in theta is especially prominent at frontal channels, whereas a difference in alpha power is prominent at the

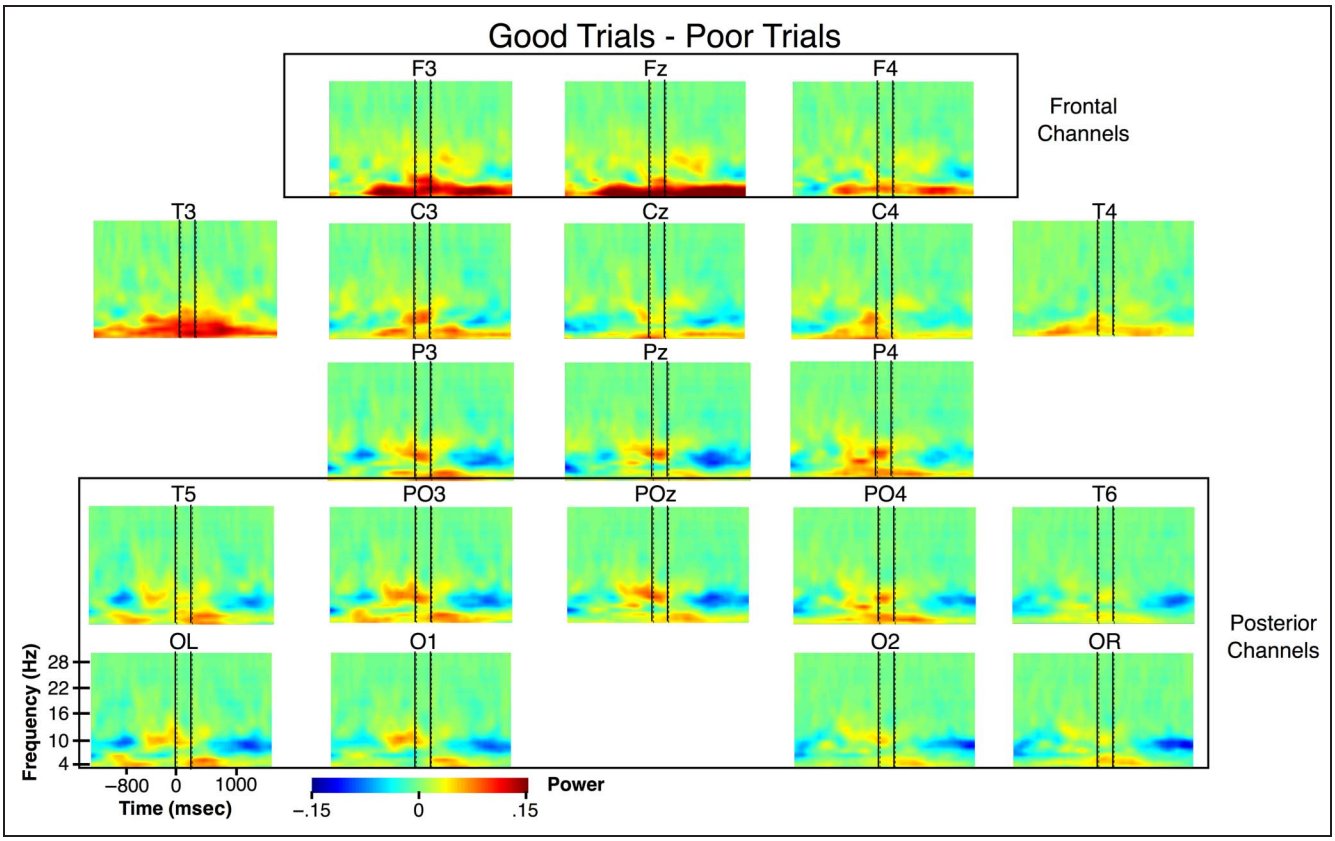

Figure 7. Spectrogram for good trials minus poor trials at all electrode sites measured. Each spectrogram represents spectral power at all frequencies from 4 to 30 Hz for each of the sites measured during Experiment 2.

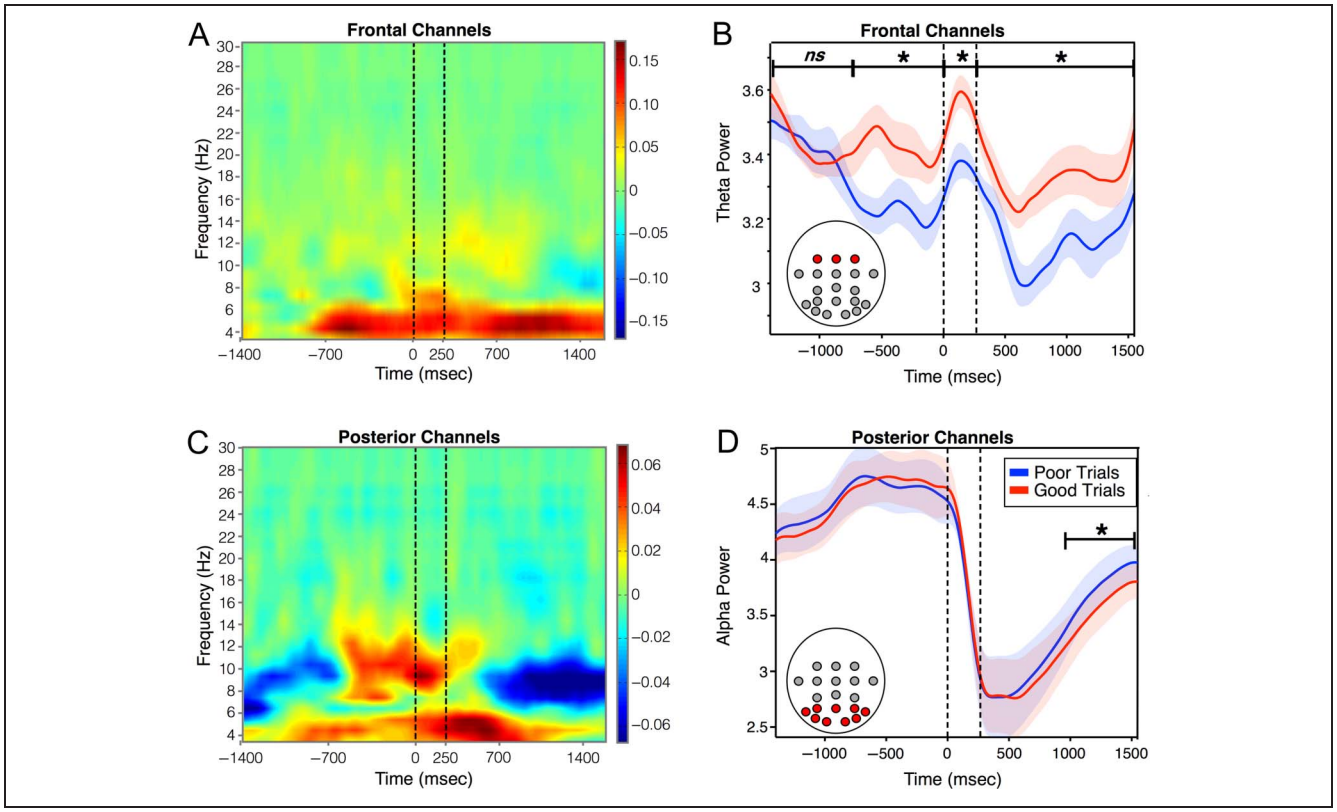

Figure 8. Theta and alpha power as a function of whole-report performance. (A) Spectral power for good trials minus poor trials at frontal electrode sites. (B) Mean theta power (4–7 Hz) as a function of trial performance. (C) Spectral power for good trials minus poor trials at posterior electrode sites. (D) Mean alpha power (8–12 Hz) as a function of trial performance.

posterior channels. Collapsing across frontal electrodes (Figure 8A), we found that decreased frontal theta power (4–7 Hz) predicted poor performance trials (Figure 8B) starting about midway through the pretrial period and sustaining throughout the retention interval. To summarize significance for key trial events, we examined theta power for early and late pretrial periods, the encoding period, and the retention interval (Figure 8B). Time-point zero corresponds to stimulus onset. We found that theta did not predict performance in the first half of the pre-trial period (−1400 to −700 msec), $t(22) = 0.845$, $p = .80$, but began to predict performance in the second half of the pretrial period, $t(22) = 2.05$, $p = .026$. Theta power continued to predict working memory performance throughout the encoding period (0–250 msec), $t(22) = 3.06$, $p < .01$, and during the retention interval (250–1550 msec), $t(22) = 2.80$, $p < .01$.[1]

Collapsing across posterior channels (Figure 8C), we found that decreased alpha power (8–12 Hz) was associated with higher performance, starting near the end of the retention interval (around 800 msec). We again determined alpha's ability to predict trial performance during key trial periods (Figure 8D). Unlike frontal theta, posterior alpha power did not predict performance in the pretrial period (−1400 to 0 msec), $t(22) = 0.62$, $p = .73$, or in the encoding period (0 to 250 msec), $t(22) = 1.16$, $p = .13$. However, decreased alpha power started to predict better performance during the retention interval (250–1550 msec), $t(22) = 2.50$, $p = .01$. This effect was driven by the second half of the retention interval. Alpha power was no different for good and poor trials in the first half of the retention interval (250–950 msec), $t(22) = 1.2$, $p = .12$, but was significantly modulated in the second half of the retention interval, $t(22) = 2.57$, $p < .01$, perhaps indicating that participants were less likely to drop items from memory toward the end of successful trials. This finding is consistent with previous work showing that greater alpha-band desynchronization is associated with increasing cognitive load.

DISCUSSION

We investigated two plausible models for how within-participant variance in working memory performance within a session gives rise to individual differences in working memory capacity. To do this, we developed and validated a novel whole-report task that provides a trial-by-trial estimate of working memory successes and failures. Examining several criteria, we found evidence against a coarse lapse model and substantial positive evidence in favor of an attentional control model. First, we found that failure rates increased as the task-load increased. This result is inconsistent with lapse models that assume that such failures should be equivalent irrespective

of task demands (Rouder et al., 2008, 2011) but is consistent with an attentional control account (Fukuda, Woodman, & Vogel, in press). Second, we observed that, although low-capacity individuals had more complete performance failures than high-capacity individuals, this higher failure rate was better explained as a downward shift of performance distributions for the low-capacity individuals. We found no evidence for a bimodal distribution of performance, as would be predicted by a lapse model. This clear distinction between the two hypothesized performance distributions was confirmed by using simulations to test the lapse and attentional control models. We found that, although both models could simulate mean performance levels, only the attentional control model produced the distribution of outcomes observed in the data. Finally, our neural data indicate that performance failures are associated with changes in oscillatory signatures of attentional control: decreased frontal theta power and increased posterior alpha power.

In addition to providing evidence for the attentional control model, our data allowed us to test several potential mechanisms that potentially underlie performance failures. One plausible explanation of individual differences in the rate of performance failures is the effect of time within the session; individual differences could simply be due to how quickly participants learn the task (e.g., many failures at beginning) or become fatigued (e.g., many failures at the end). This hypothesis predicts that individual differences are disproportionately explained by performance at the very beginning or end of the experiment. Contrary to this hypothesis we find that performance failures occur consistently throughout the experiment and that the differences between high and low-capacity individuals are stable throughout the entire experiment. A second explanation for performance failures is simple task noncompliance. Although we instruct participants to keep their eyes open and focused on the central fixation dot, participants blinking or moving their eyes away from fixation could result in poor performance. However, we found no relationship between the rate of EOG artifacts and performance failures and that these failure rates are preserved even after excluding trials with ocular artifacts. A third reason for poor performance is that there is insufficient sensory encoding of the memory array items. However, we found no difference in the visual evoked response (P1 and N1) to the memory array items between poor and good performance trials. Together, these results suggest that these working memory performance failures are not simply due to slow learning, fatigue, ocular artifacts, or poor sensory encoding.

In contrast to the above results, our neural measures provide positive evidence that performance failures are related to well-known oscillatory markers of attentional control mechanisms: frontal theta and posterior alpha. We find that mean frontal theta power is higher on success trials than for failure trials and that this difference begins a few hundred milliseconds before the trial even begins and persists throughout the retention period of the task. There is a substantial literature relating frontal theta power to attentional control and memory success, and the current work is broadly consistent with these findings. In addition, the finding that theta power can distinguish between working memory successes and failures before the trial has begun suggests that it reflects preparatory mechanisms of attentional control that need to be engaged in advance to adequately perform these tasks (Leber, Turk-Browne, & Chun, 2008). Likewise, toward the end of the retention interval, we observe that increased posterior alpha power predicts performance failures. This may reflect an inability to sustain the alpha desynchronization that is necessary for ongoing memory storage. Together, our findings fit well within the literature showing that increased theta and decreased alpha at encoding predict successful memory performance (Stipacek et al., 2003; Klimesch, 1999; Klimesch et al., 1997). Our findings are also consistent with hypotheses about (1) the strong relationship between working memory and attention (Unsworth et al., 2014; Chun, Golomb, & Turk-Browne, 2011; Engle & Kane, 2004), (2) the trial-to-trial variability of attention (Esterman, Rosenberg, & Noonan, 2014; Esterman, Noonan, Rosenberg, & DeGutis, 2013), and (3) the importance of prefrontal networks in sustaining attentional control (Liesefeld, Liesefeld, & Zimmer, 2014; Giesbrecht, Woldorff, Song, & Mangun, 2003).

Individual differences in visual working memory capacity are robust, stable, and predictive of fluid intelligence and have been proposed to be due to variations in attentional control (Unsworth et al., 2014; Fukuda et al., 2010; Engle et al., 1999). However, a compelling alternative model proposes that these differences are instead due to how frequently the individual is completely disengaged from the task at hand. Our current results reject such a coarse lapse model and suggest that graded fluctuations in attentional control from trial to trial within a session drive the individual differences in capacity that are observed in traditional aggregate measures of performance. Failed attentional control on a trial would be expected to produce a wide swath of processing errors such as insufficient individuation, poor resolution, item position swapping, and retrieval failures. The present work suggests that the ability to prevent such failures by consistently engaging attentional control mechanisms during challenging tasks is a central component of an individual's cognitive ability.

Acknowledgments

This study was supported by grants NIH-R01-MH087214 and Office of Naval Research N00014-12-1-0972. Data from this paper are available for download through Open Science Framework. Go to the Open Science Framework Web site and search for the title of this paper.

Reprint requests should be sent to Kirsten C. S. Adam, Department of Psychology, 1227 University of Oregon, Eugene, OR 97401, or via e-mail: kadam@uoregon.edu.

Note

1. p values are not corrected for multiple comparisons; the Bonferroni-corrected threshold for four comparisons is $p = .013$. The pretrial effect ($p = .026$) does not survive this very conservative thresholding. If we choose only a single time window (−700 to 0 msec), we can check for this effect at all electrodes of interest (see Figure 7): Fz ($p = .026$), F3 ($p = .037$), F4 ($p = .1$), T3 ($p = .03$).

REFERENCES

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436.

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences, 18,* 414–421.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology, 62,* 73–101.

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

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

Coyle, T. (2003). A review of the worst performance rule: Evidence, theory, and alternative hypotheses. *Intelligence, 31,* 567–587.

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

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language, 25,* 1–18.

Deiber, M.-P., Missonnier, P., Bertrand, O., Gold, G., Fazio-Costa, L., Ibañez, V., et al. (2007). Distinction between perceptual and attentional processing in working memory tasks: A study of phase-locked and induced oscillatory brain dynamics. *Journal of Cognitive Neuroscience, 19,* 158–172.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134,* 9–21.

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology, 30,* 257–303.

Duncan, J., Schramm, M., Thompson, R., & Dumontheil, I. (2012). Task rules, working memory, and fluid intelligence. *Psychonomic Bulletin & Review, 19,* 864–870.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross, (Ed.), *Psychology of learning and motivation* (pp. 145–199). NY: Elsevier.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128,* 309–331.

Esterman, M., Noonan, S. K., Rosenberg, M., & DeGutis, J. (2013). In the zone or zoning out? Tracking behavioral and neural fluctuations during sustained attention. *Cerebral Cortex, 23,* 2712–2723.

Esterman, M., Rosenberg, M. D., & Noonan, S. K. (2014). Intrinsic fluctuations in sustained attention and distractor processing. *Journal of Neuroscience, 34,* 1724–1730.

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review, 17,* 673–679.

Fukuda, K., & Vogel, E. K. (2009). Human variation in overriding attentional capture. *Journal of Neuroscience, 29,* 8726–8733.

Fukuda, K., & Vogel, E. K. (2011). Individual differences in recovery time from attentional capture. *Psychological Science, 22,* 361–368.

Fukuda, K., Woodman, G. F., & Vogel, E. K. (in press). Individual differences in visual working memory capacity: Contributions of attentional control to storage. In *Attention & performance XXV*. Elsevier.

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

Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex, 7,* 374–385.

Giesbrecht, B., Woldorff, M., Song, A., & Mangun, G. (2003). Neural mechanisms of top–down control during spatial and feature attention. *Neuroimage, 19,* 496–512.

Hsieh, L.-T., & Ranganath, C. (2014). Frontal midline theta oscillations during working memory maintenance and episodic encoding and retrieval. *Neuroimage, 85,* 721–729.

Itthipuripat, S., Wessel, J. R., & Aron, A. R. (2013). Frontal theta is a signature of successful working memory manipulation. *Experimental Brain Research, 224,* 255–262.

Jensen, O., & Tesche, C. D. (2002). Frontal theta activity in humans increases with memory load in a working memory task. *European Journal of Neuroscience, 15,* 1395–1399.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Research Reviews, 29,* 169–195.

Klimesch, W., Doppelmayr, M., Schimke, H., & Ripper, B. (1997). Theta synchronization and alpha desynchronization in a memory task. *Psychophysiology, 34,* 169–176.

Leber, A. B., Turk-Browne, N. B., & Chun, M. M. (2008). Neural predictors of moment-to-moment fluctuations in cognitive flexibility. *Proceedings of the National Academy of Sciences, U.S.A., 105,* 13592–13597.

Liesefeld, A. M., Liesefeld, H. R., & Zimmer, H. D. (2014). Intercommunication between prefrontal and posterior brain regions for protecting visual working memory from distractor interference. *Psychological Science, 25,* 325–333.

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

McCollough, A. W., Machizawa, M. G., & Vogel, E. K. (2007). Electrophysiological measures of maintaining representations in visual working memory. *Cortex, 43,* 77–94.

McVay, J. C., & Kane, M. J. (2010). Does mind wandering reflect executive function or executive failure? Comment on Smallwood and Schooler (2006) and Watkins (2008). *Psychological Bulletin, 136,* 188–197; discussion 198–207.

Morey, R. D. (2011). A Bayesian hierarchical model for the measurement of working memory capacity. *Journal of Mathematical Psychology, 55,* 8–24.

Mrazek, M. D., Smallwood, J., Franklin, M. S., Chin, J. M., Baird, B., & Schooler, J. W. (2012). The role of mind-wandering in measurements of general aptitude. *Journal of Experimental Psychology: General, 141,* 788–798.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442.

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, U.S.A., 105,* 5975–5979.
Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin & Review, 18,* 324–330.
Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review, 119,* 807–830.
Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological Bulletin, 132,* 946–958.
Stipacek, A., Grabner, R. H., Neuper, C., Fink, A., & Neubauer, A. C. (2003). Sensitivity of human EEG alpha band desynchronization to different working memory components and increasing levels of memory load. *Neuroscience Letters, 353,* 193–196.
Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28,* 127–154.
Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology, 71,* 1–26.
Van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121,* 124–149.
Weissman, D. H., Roberts, K. C., Visscher, K. M., & Woldorff, M. G. (2006). The neural bases of momentary lapses in attention. *Nature Neuroscience, 9,* 971–978.
Zakrzewska, M. Z., & Brzezicka, A. (2014). Working memory capacity as a moderator of load-related frontal midline theta variability in Sternberg task. *Frontiers in Human Neuroscience, 8*, Article 399.