

# The reliability and stability of visual working memory capacity

Z. Xu<sup>1</sup> · K. C. S. Adam<sup>2</sup> · X. Fang<sup>1</sup> · E. K. Vogel<sup>2</sup>

Published online: 7 April 2017  
© Psychonomic Society, Inc. 2017

**Abstract** Because of the central role of working memory capacity in cognition, many studies have used short measures of working memory capacity to examine its relationship to other domains. Here, we measured the reliability and stability of visual working memory capacity, measured using a single-probe change detection task. In Experiment 1, the participants ( $N = 135$ ) completed a large number of trials of a change detection task (540 in total, 180 each of set sizes 4, 6, and 8). With large numbers of both trials and participants, reliability estimates were high ( $\alpha > .9$ ). We then used an iterative down-sampling procedure to create a look-up table for expected reliability in experiments with small sample sizes. In Experiment 2, the participants ( $N = 79$ ) completed 31 sessions of single-probe change detection. The first 30 sessions took place over 30 consecutive days, and the last session took place 30 days later. This unprecedented number of sessions allowed us to examine the effects of practice on stability and internal reliability. Even after much practice, individual differences were stable over time (average between-session  $r = .76$ ).

**Keywords** Visual working memory · Reliability · Change detection

Working memory (WM) capacity is a core cognitive ability that predicts performance across many domains. For example, WM capacity predicts attentional control, fluid intelligence,

and real-world outcomes such as perceiving hazards while driving (Engle, Tuholski, Laughlin, & Conway, 1999; Fukuda, Vogel, Mayr, & Awh, 2010; Wood, Hartley, Furley, & Wilson, 2016). For these reasons, researchers are often interested in devising brief measures of WM capacity to investigate the relationship of WM capacity to other cognitive processes. However, truncated versions of WM capacity tasks could potentially be inadequate for reliably measuring an individual's capacity. Inadequate measurement could obscure correlations between measures, or even differences in performance between experimental conditions. Furthermore, although WM capacity is considered to be a stable trait of the observer, little work has directly examined the role of extensive practice in the measurement of WM capacity over time. This is of particular concern because of the popularity of research examining whether training affects WM capacity (Melby-Lervåg & Hulme, 2013; Shipstead, Redick, & Engle, 2012). Extensive practice on any given cognitive task has the potential to significantly alter the nature of the variance that determines performance. For example, extensive practice has the potential to induce a restriction-of-range problem, in which the bulk of the observers reach similar performance levels—thus reducing any opportunity to observe correlations with other measures. Consequently, a systematic study of the reliability and stability of WM capacity measures is critical for improving the measurement and reproducibility of major phenomena in this field.

In the present study, we sought to establish the reliability and stability of one particular WM capacity measure: change detection. Change detection measures of visual WM have gained popularity as a means of assessing individual differences in capacity. In a typical change detection task, participants briefly view an array of simple visual items (for ~100–500 ms), such as colored squares, and remember these items across a short delay (~1–2 s). At test, observers are presented

---

Z. Xu and K. C. S. Adam contributed equally to this work.

✉ K. C. S. Adam  
kadam1@uchicago.edu

<sup>1</sup> School of Psychology, Southwest University, Chongqing, China

<sup>2</sup> Department of Psychology, University of Chicago, 940 E 57th Street, Chicago, Illinois 60637, USA

with an item at one of the remembered locations, and they indicate whether the presented test item is the same as the remembered item (“no-change” trial) or is different (“change trial”). Performance can be quantified as raw accuracy or converted into a capacity estimate (“ $K$ ”). In capacity estimates, performance for change trials and no-change trials is calculated separately as hits (the proportion of correct change trials) and false alarms (the proportion of incorrect no-change trials) and converted into a set-size-dependent score (Cowan, 2001; Pashler, 1988; Rouder, Morey, Morey, & Cowan, 2011).

Several beneficial features of change detection tasks have led to their increased popularity. First, change detection memory tasks are simple and short enough to be used with developmental and clinical populations (e.g., Cowan, Fristoe, Elliott, Brunner, & Saults, 2006; Gold, Wilk, McMahon, Buchanan, & Luck, 2003; Lee et al., 2010). Second, the relatively short length of trials lends the task well to neural measures that require large numbers of trials. In particular, neural studies employing change detection tasks have provided strong corroborating evidence of capacity limits in WM (Todd & Marois, 2004; Vogel & Machizawa, 2004) and have yielded insights into potential mechanisms underlying individual differences in WM capacity (for a review, see Luria, Balaban, Awh, & Vogel, 2016). Finally, change detection tasks and closely related memory-guided saccade tasks can be used with animal models from pigeons (Gibson, Wasserman, & Luck, 2011) to nonhuman primates (Buschman, Siegel, Roy, & Miller, 2011), providing a rare opportunity to directly compare behavior and neural correlates of task performance across species (Elmore, Magnotti, Katz, & Wright, 2012; Reinhart et al., 2012).

A main aim of this study is to quantify the effect of measurement error and sample size on the reliability of change detection estimates. In previous studies, change detection estimates of capacity have yielded good reliability estimates (e.g., Pailian & Halberda, 2015; Unsworth, Fukuda, Awh, & Vogel, 2014). However, measurement error can vary dramatically with the number of trials in a task, thus impacting reliability; Pailian and Halberda found that reliability of change detection estimates greatly improved when the number of trials was increased. Researchers frequently employ vastly different numbers of trials and participants in studies of individual differences, but the effect of trial number on change detection reliability has never been fully characterized. In studies using large batteries of tasks, time and measurement error are forces working in opposition to one another. When researchers want to minimize the amount of time that a task takes, measures are often truncated to expedite administration. Such truncated measures increase measurement noise and potentially harm the reliability of the measure. At present, there is no clear understanding of the minimum number of either participants or trials that is necessary to obtain reliable estimates of change detection capacity.

In addition to measurement error within-session, reliability of individual differences could be compromised with extensive practice. Previously, it was found that visual WM capacity estimates were stable ( $r = .77$ ) after 1.5 years between testing sessions (Johnson et al., 2013). However, the effect of extensive practice on change detection estimates of capacity has yet to be characterized. Extensive practice could harm the reliability and stability of measures in a couple of ways. First, it is possible that participants could improve so much that they reach performance ceiling, thus eliminating variability between individuals. Second, if individual differences are due to the utilization of optimal versus suboptimal strategies, then participants might converge to a common mean after engaging in extensive practice and finding optimal task strategies. Both of these hypothetical possibilities would call into question the true stability of WM capacity estimates, and likewise severely harm the statistical reliability of the measure. As such, in Experiment 2 we directly quantified the extent of extensive practice on the stability of WM capacity estimates.

## Overview of experiments

We measured the reliability and stability of a single-probe change detection measure of visual WM capacity. In Experiment 1, we measured the reliability of capacity estimates obtained with a commonly used version of the color change detection task for a relatively large number of participants ( $n = 135$ ) and a larger than typical number of trials ( $t = 540$ ). In Experiment 2, we measured the stability of capacity estimates across an unprecedented number of testing sessions (31). Because of the large number of sessions, we could investigate the stability of change detection estimates after extended practice and over a period of 60 days.

## Experiment 1

### Materials and method

**Participants** A total of 137 individuals (102 females, 35 males; mean age = 19.97,  $SD = 1.07$ ) with normal or corrected-to-normal vision participated in the experiment. Participants provided written informed consent, and the study was approved by the Ethics Committee at Southwest University. Participants received monetary compensation for their participation. Two participants were excluded because they had negative average capacity values, resulting in a final sample of 135 participants.

**Stimuli** The stimuli were presented on monitors with a refresh rate of 75 Hz and a screen resolution of  $1,024 \times 768$ . Participants sat approximately 60 cm from the screen, though

a chinrest was not used so all visual angle estimates are approximate. In addition, there were some small variations in monitor size (five 16-in. CRT monitors, three 19-in. LCD monitors) in testing rooms, leading to small variations in the size of the colored squares from monitor to monitor. Details are provided about the approximate range in degrees of visual angle.

All stimuli were generated in MATLAB (The MathWorks, Natick, MA) using the Psychophysics Toolbox (Brainard, 1997). Colored squares (51 pixels; range of 1.55° to 2.0° visual angle) served as memoranda. Squares could appear anywhere within an area of the monitor subtending approximately 10.3° to 13.35° horizontally and 7.9° to 9.8° vertically. Squares could appear in any of nine distinct colors, and colors were sampled without replacement within each trial (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; orange = 255 128 0; white = 255 255 255; black = 0 0 0). Participants were instructed to fixate a small black dot (approximate range: .36° to .47° of visual angle) at the center of the display.

**Procedures** Each trial began with a blank fixation period of 1,000 ms. Then, participants briefly viewed an array of four, six, or eight colored squares (150 ms), which they remembered across a blank delay period (1,000 ms). At test, one colored square was presented at one of the remembered locations. The probabilities were equal that the probed square was the same color (no-change trial) or was a different color (change trial). Participants made an unspeeded response by pressing the “z” key, if the color was the same, or the “/” key, if the color was different. Participants completed 180 trials of set sizes 4, 6, and 8 (540 trials total). Trials were divided into nine blocks, and participants were given a brief rest period (30 s) after each block. To calculate capacity, change detection accuracy was transformed into a  $K$  estimate using Cowan’s (2001) formula  $K = N \times (H - FA)$ , where  $N$  represents the set size,  $H$  is the hit rate (proportion of correct responses to change trials), and  $FA$  is the false alarm rate (proportion of incorrect responses to no-change trials). Cowan’s formula is best for single-probe displays like the one employed here. For change detection tasks using whole-display probes, Pashler’s (1988) formula may be more appropriate (Rouder et al., 2011).

## Results

Descriptive statistics for each set size condition are shown in Table 1, and data for both Experiments 1 and 2 are available online at the website of the Open Science Framework, at <https://osf.io/g7txf/>. We observed a significant difference in performance across set sizes,  $F(2, 268) = 20.6, p < .001, \eta_p^2$

**Table 1** Descriptive statistics for Experiment 1

	Mean $K$	$SD$	Min	Max	Kurtosis	Skewness
Set Size 4	2.32	0.70	0.58	3.87	−0.49	−0.35
Set Size 6	2.10	0.97	0.07	4.80	−0.18	0.34
Set Size 8	1.99	0.97	−0.18	4.53	−0.52	0.14
Average	2.14	0.82	0.38	4.31	−0.47	0.07

Descriptive statistics are shown separately for each set size and for the average of the three set sizes. Kurtosis and skewness values are both centered around 0. Neither kurtosis nor skewness was credibly nonnormal in any condition (Cramer, 1997).

= .133, and polynomial contrasts revealed a significant linear trend,  $F(1, 134) = 36.48, p < .001, \eta_p^2 = .214$ , indicating that the average performance declined slightly with increased memory load.

**Reliability of the full sample: Cronbach’s alpha** We computed Cronbach’s alpha (unstandardized) using  $K$  scores from the three set sizes as items (180 trials contributing to each item), and obtained a value of  $\alpha = .91$  (Cronbach, 1951). We also computed Cronbach’s alpha using  $K$  scores from the nine blocks of trials (60 trials contributing to each item) and obtained a nearly identical value of  $\alpha = .92$ . Finally, we computed Cronbach’s alpha using raw accuracy for single trials (540 items), and obtained an identical value of  $\alpha = .92$ . Thus, change detection estimates had high internal reliability for this large sample of participants, and the precise method used to divide trials into “items” does not impact Cronbach’s alpha estimates of reliability for the full sample. Furthermore, using raw accuracy versus bias-corrected  $K$  scores did not impact reliability.

**Reliability of the full sample: Split-half** The split-half correlation of the  $K$  scores for even and odd trials was reliable,  $r = .88, p < .001, 95\% \text{ CI } [.84, .91]$ . Correcting for attenuation yielded a split-half correlation value of  $r = .94$  (Brown, 1910; Spearman, 1910). Likewise, the capacity scores from individual set sizes correlated with each other:  $r_{ss4-ss6} = .84, p < .001, 95\% \text{ CI } [.78, .88]$ ;  $r_{ss6-ss8} = .79, p < .001, 95\% \text{ CI } [.72, .85]$ ;  $r_{ss4-ss8} = .76, p < .001, 95\% \text{ CI } [.68, .83]$ . Split-half correlations for individual set sizes yielded Spearman–Brown-corrected correlation values of  $r = .91$  for set size 4,  $r = .86$  for set size 6, and  $r = .76$  for set size 8, respectively.

The drop in capacity from set size 4 to set size 8 has been used in the literature as a measure of filtering ability. However, the internal reliability of this difference score has typically been low (Pailian & Halberda, 2015; Unsworth et al., 2014). Likewise, we found here that the split-half reliability of the performance decline from set size 4 to set size 8 (“4–8 Drop”) was low, with a Spearman–Brown-corrected correlation value of  $r = .24$ . Although weak, this correlation is of the same

strength that was reported in earlier work (Unsworth et al., 2014). The split-half reliability of the performance decline from set size 4 to set size 6 was slightly higher,  $r = .39$ , and the split-half reliability of the difference between set size 6 and set size 8 performance was very low,  $r = .08$ . The reliability of differences scores can be impacted both by (1) the internal reliability of each measure used to compute the difference and (2) the degree of correlation between the two measures (Rodebaugh et al., 2016). Although the internal reliability of each individual set size was high, the positive correlation between set sizes may have decreased the reliability of the set size difference scores.

**An iterative down-sampling approach** To investigate the effects of sample size and trial number on the reliability estimates, we used an iterative down-sampling procedure. Two reliability metrics were assessed: (1) Cronbach's alpha using single-trial accuracy as items, and (2) split-half correlations using all trials. For the down-sampling procedure, we randomly sampled participants and trials from the full dataset. The number of participants ( $n$ ) was varied from 5 to 135 in steps of 5. The number of trials ( $t$ ) was varied from 5 to 540 in steps of 5. Number of participants and number of trials were factorially combined (2,916 cells total). For each cell in the design, we ran 100 sampling iterations. On each iteration,  $n$  participants and  $t$  trials were randomly sampled from the full dataset and reliability metrics were calculated for the sample.

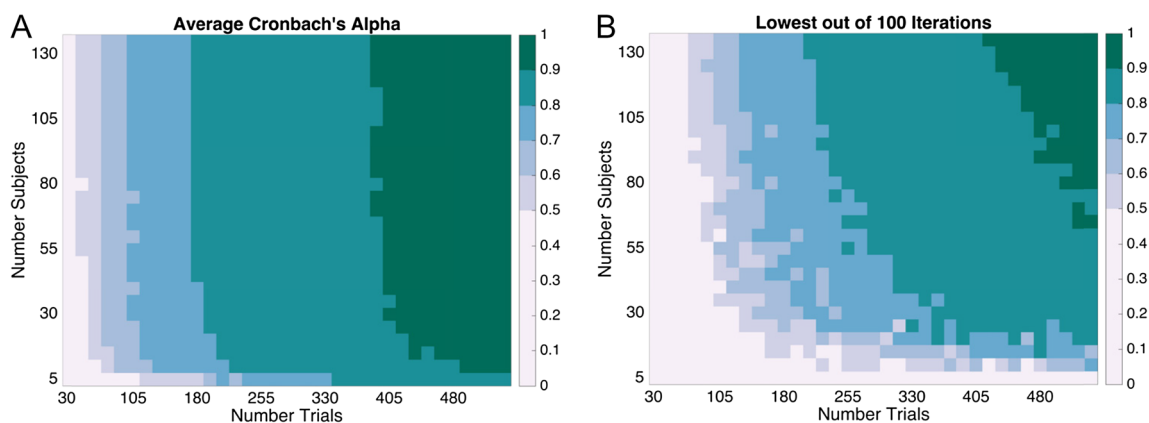
Figure 1 shows the results of the down-sampling procedure for Cronbach's alpha. Figure 2 shows the results of the down-sampling procedure for split-half reliability estimates. In each plot, we show both the average reliabilities obtained across the 100 iterations (Figs. 1a and 2a) and the worst reliabilities obtained across the 100 iterations (Figs. 1b and 2b). Conceptually, we could think of each iteration of the down-sampling procedure as akin to running one “experiment,” with participants randomly sampled from our “population” of 137. Although it is good to know the average expected reliability

across many experiments, the typical experimenter will run an experiment only once. Thus, considering the “worst case scenario” is instructive for planning the number of participants and the number of trials to be collected. For a more complete picture of the breadth of the reliabilities obtained, we can also consider the variability in reliabilities across iterations ( $SD$ ) and the range of reliability values (Fig. 2c and d). Finally, we repeated this iterative down-sampling approach for each individual set size. The average reliability as well as the variability of the reliabilities for individual set sizes are shown in Fig. 3. Note that each set size begins with 1/3 as many trials as in Figs. 1 and 2.

Next, we looked at some potential characteristics of samples with low reliability (e.g., iterations with particularly low vs. high reliability). We ran 500 sampling iterations of 30 participants and 120 trials, then we did a median split for high- versus low-reliability samples. No significant differences emerged in the mean ( $p = .86$ ), skewness ( $p = .60$ ), or kurtosis ( $p = .70$ ) values of high- versus low-reliability samples. There were, however, significant effects of sample range and variability. As would be expected, samples with higher reliability had larger standard deviations,  $t(498) = 26.7$ ,  $p < .001$ , 95% CI [.14, .17], and wider ranges,  $t(498) = 15.2$ ,  $p < .001$ , 95% CI [.52, .67], than samples with low reliability.

#### A note for fixed capacity + attention estimates of capacity

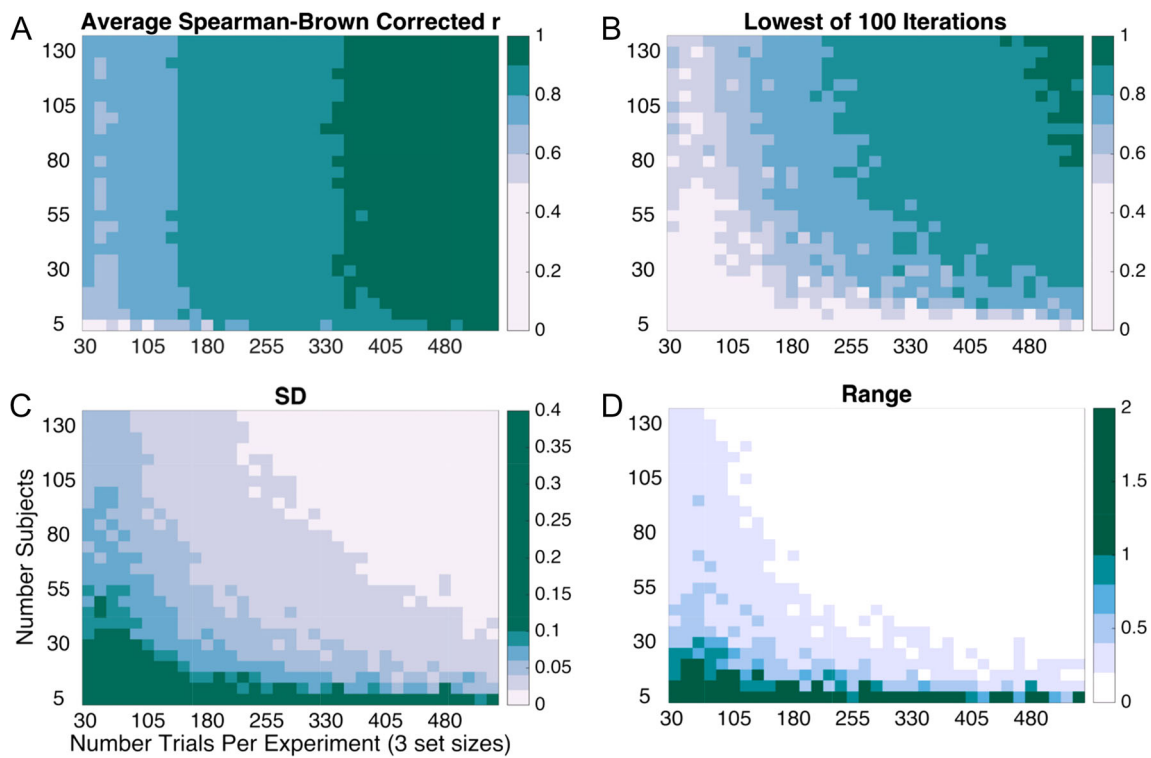
So far, we have discussed only the most commonly used methods of estimating WM capacity ( $K$  scores and percentages correct). Other methods of estimating capacity have been used, and we now briefly mention one of them. Rouder and colleagues (2008) suggested adding an attentional lapse parameter to estimates of visual WM capacity, a model referred to as fixed capacity + attention. Adding an attentional lapse parameter accounts for trials in which participants are inattentive to the task at hand. Specifically, participants commonly make errors on trials that should be well within capacity limits (e.g., set size 1), and adding a lapse parameter can help explain



**Fig. 1** Cronbach's alpha as a function of the number of trials and the number of participants in Experiment 1. In each cell, Cronbach's alpha was computed for  $t$  trials ( $x$ -axis) and  $n$  participants ( $y$ -axis). (a) Average

reliabilities across 100 iterations. (b) Minimum reliabilities obtained (worst random samples of participants and trials)



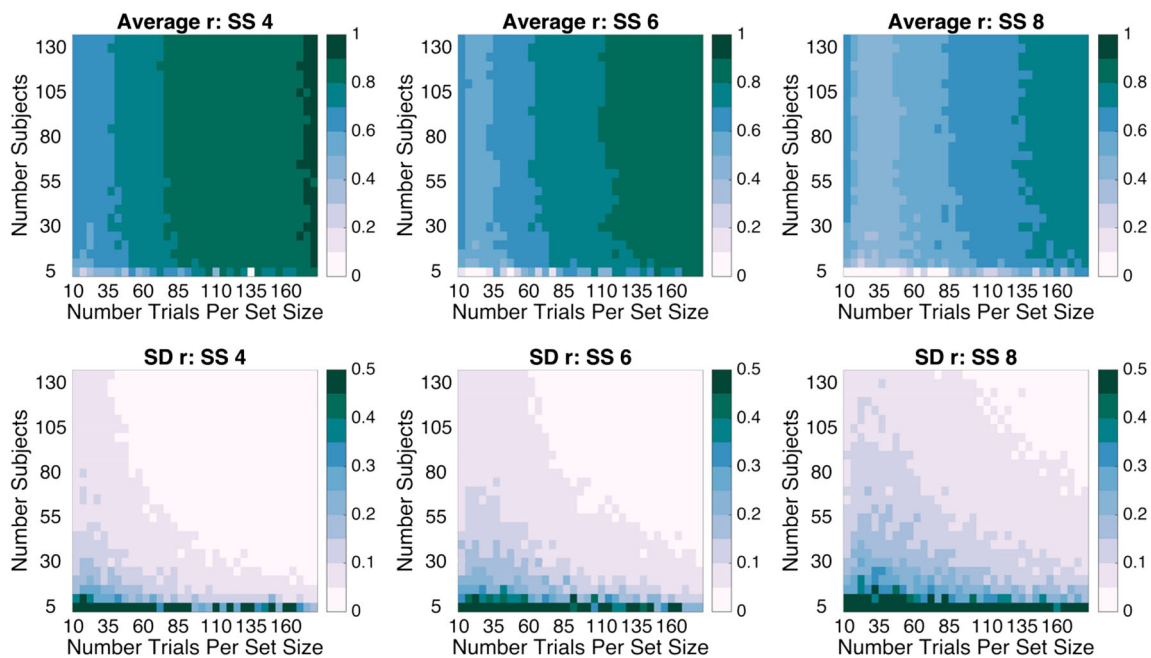


**Fig. 2** Spearman–Brown-corrected split-half reliability estimates as a function of the numbers of trials and participants in Experiment 1. (a) Average reliabilities across 100 iterations. (b) Minimum reliabilities

obtained (worst random samples of participants and trials). (c) Standard deviations of the reliabilities obtained across samples. (d) Range of reliability values obtained across samples

these anomalous dips in performance. Unlike typical estimates of capacity, in which a  $K$  value is computed directly for performance for each set size and then averaged, this model uses a log-likelihood estimation technique that estimates a single

capacity parameter by simultaneously considering performance across all set sizes and/or change probability conditions. Critically, this model assumes that data are obtained for at least one subcapacity set size, and that any error made



**Fig. 3** Spearman–Brown-corrected split-half reliability estimates for each set size in Experiment 1. Top panels: Average reliabilities for each set size. Bottom panels: Standard deviations of the reliabilities for each set size across 100 down-sampling iterations

on this set size reflects an attentional lapse. If the model is fit to data that lack at least one subcapacity set size (e.g., one or two items), then the model will fit poorly and provide nonsensical parameter estimates.

Recently, Van Snellenberg, Conway, Spicer, Read, and Smith (2014) used the fixed capacity + attention model to calculate capacity for a change detection task, and they found that the reliability of the model's capacity parameter was low ( $r = .32$ ) and did not correlate with other WM tasks. Critically, however, this study used only relatively high set sizes (4 and 8) and lacked a subcapacity set size, so model fits were likely poor. Using code made available by Rouder et al., we fit a fixed capacity + attention model to our data (Rouder, n.d.). We found that when this model is misapplied (i.e., used on data without at least one subcapacity set size), the internal reliability of the capacity parameter was low ( $r$  uncorrected = .35) and was *negatively* correlated with raw change detection accuracy,  $r = -.25$ ,  $p = .004$ . If we had only applied this model to our data, we would have mistakenly concluded that change detection measures offer poor reliability and do not correlate with other measures of WM capacity.

## Discussion

Here, we have shown that when sufficient numbers of trials and participants are collected, the reliability of change detection capacity is remarkably high ( $r > .9$ ). On the other hand, a systematic down-sampling method revealed that insufficient trials or insufficient participant numbers could dramatically reduce the reliability obtained in a single experiment. If researchers hope to measure the correlation between visual WM capacity and some other measure, Figs. 1 and 2 can serve as an approximate guide to expected reliability. Because we had only a single sample of the largest  $n$  (137), we cannot make definitive claims about the reliabilities of future samples of this size. However, given the stabilization of correlation coefficients with large sample sizes and the extremely high correlation coefficient obtained, we can be relatively confident that the reliability estimate for our full sample ( $n = 137$ ) would not change substantially in future samples of university students. Furthermore, we can make claims about how the reliability of small, well-defined subsamples of this “population” can systematically deviate from an empirical upper bound.

The average capacity obtained for this sample was slightly lower than some other values in the literature, typically cited as around three or four items. The slightly lower average for this sample could potentially cause some concern about the generalizability of these reliability values for future samples. For the present study's sample, the average  $K$  scores for set sizes 4 and 8 were  $K = 2.3$  and 2.0, respectively. The largest, most comparable sample to the present sample is a 495-participant sample in a work by Fukuda, Woodman, and Vogel (2015). The average  $K$  scores for set sizes 4 and 8 were

$K = 2.7$  and 2.4, respectively, and the task design was nearly identical (150-ms encoding time, 1,000-ms retention interval, no color repetitions allowed, and set sizes 4 and 8). The difference of 0.3–0.4 items between these two samples is relatively small, though likely significant. However, for the purposes of estimating reliability, the variance of the distribution is more important than the mean. The variabilities observed in the present sample ( $SD = 0.7$  for set size 4,  $SD = 0.97$  for set size 8) were very similar to those observed in the Fukuda et al. sample ( $SD = 0.6$  for set size 4 and  $SD = 1.2$  for set size 8), though unfortunately the Fukuda et al. study did not report reliability. Because of the nearly identical variabilities of scores across these two samples, we can infer that our reliability results would indeed generalize to other large samples for which change detection scores have been obtained.

We recommend applying an iterative down-sampling approach to other measures when expediency of task administration is valued, but reliability is paramount. The stats-savvy reader may note that the Spearman–Brown prophecy formula also allows one to calculate how many observations must be added to improve the expected reliability, according to the formula

$$N = \frac{\rho_{xx'}^* (1 - \rho_{xx'})}{\rho_{xx'} (1 - \rho_{xx'}^*)}$$

where  $\rho_{xx'}^*$  is the desired correlation strength,  $\rho_{xx'}$  is the observed correlation, and  $N$  is the number of times that a test length must be multiplied to achieve the desired correlation strength. Critically, however, this formula *does not account for the accuracy of the observed correlation*. Thus, if one starts from an unreliable correlation coefficient obtained with a small number of participants and trials, one will obtain an unreliable estimate of the number of observations needed to improve the correlation strength. In experiments such as this one, both the number of trials and the number of participants will drastically change estimates of the number of participants needed to observe correlations of a desired strength.

Let's take an example from our iterative down-sampling procedure. Imagine that we ran 100 experiments, each with 15 participants and 150 total trials of change detection. Doing so, we would obtain 100 different estimates of the strength of the true split-half correlation. We could then apply the Spearman–Brown formula to each of these 100 estimates in order to calculate the number of trials needed to obtain a desired reliability of  $r = .8$ . So doing, we would find that, on average, we would need around 140 trials to obtain the desired reliability. However, because of the large variability in the observed correlation strength ( $r = .37$  to  $.97$ ), if we had only run the “best case” experiment ( $r = .97$ ), we would estimate that we need only 18 trials to obtain our desired reliability of  $r = .8$  with 15 participants. On the other hand, if we had run the

“worst case” experiment ( $r = .37$ ), then we would estimate that we need 1,030 trials. There are downsides to both types of estimation errors. Although a pessimistic estimate of the number of trials needed ( $>1,000$ ) would certainly ensure adequate reliability, this might come at the cost of time and participants’ frustration. Conversely, an overly optimistic estimate of the number of trials needed ( $<20$ ) would lead to underpowered studies that would waste time and funds.

Finally, we investigated an alternative parameterization of capacity based on a model that assumes a fixed capacity and an attention lapse parameter (Rouder et al., 2008). Critically, this model attempts to explain errors for set sizes that are well within capacity limits (e.g., one item). If researchers inappropriately apply this model to change detection data with only large set sizes, they would erroneously conclude that change detection tasks yield poor reliability and fail to correlate with other estimates of capacity (e.g., Van Snellenberg et al., 2014).

In Experiment 2, we shifted our focus to the stability of change detection estimates. That is, how consistent are estimates of capacity from day to day? We collected an unprecedented number of sessions of change detection performance (31) spanning 60 days. We examined the stability of capacity estimates, defined as the correlation between individuals’ capacity estimates from one day to the next. Since capacity is thought to be a stable trait of the individual, we predicted that individual differences in capacity should be reliable across many testing sessions.

## Experiment 2

### Materials and methods

**Participants** A group of 79 individuals (22 males, 57 females; mean age = 22.67 years,  $SD = 2.31$ ) with normal or corrected-to-normal vision participated for monetary compensation. The study was approved by the Ethics Committee of Southwest University.

**Stimuli** Some experimental sessions were completed in the lab and others were completed in participants’ homes. In the lab, stimuli were presented on monitors with a refresh rate of 75 Hz. At home, stimuli were presented on laptop screens with somewhat variable refresh rates and sizes. In both cases, participants sat approximately 60 cm from the screen, though a chinrest was not used, so all visual angle estimates are approximate. In the lab there were some small variations in monitor size (five 18.5-in. LCD monitors, one 19-in. LCD monitor) in the testing rooms, leading to small variations in the sizes of the colored squares. Details are provided about the approximate range in degrees of visual angle in the lab.

All stimuli were generated in MATLAB (The MathWorks, Natick, MA) using the Psychophysics Toolbox (Brainard, 1997). Colored squares (51 pixels; range of  $1.28^\circ$  to  $1.46^\circ$  visual angle) served as the memoranda. Squares could appear anywhere within an area of the monitor subtending approximately  $14.4^\circ$  to  $14.8^\circ$  horizontally and  $8.1^\circ$  to  $8.4^\circ$  vertically. Squares could appear in any of nine distinct 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; orange = 255 128 0; white = 255 255 255; black = 0 0 0). Colors were sampled without replacement for set size 4 and set size 6 trials. Each color could be repeated up to one time in set size 8 trials (i.e., colors were sampled from a list of 18 colors, with each of the nine unique colors appearing twice). Participants were instructed to fixate a small black dot ( $\sim 0.3^\circ$  visual angle) at the center of the display.

**Procedures** Trial procedures for the change detection task were identical to Experiment 1. Participants completed a total of 31 sessions of the change detection task. In each session, participants completed a total of 120 trials (split over five blocks). There were 40 trials each of set sizes 4, 6, and 8. Participants were asked to finish the change detection task once a day for 30 consecutive days. They could do this task on their own computers or on the experimenters’ computers throughout the day. Participants were instructed that they should complete the task in a relatively quiet environment and not do anything else (e.g., talking to others) at the same time. Experimenters reminded the participants to finish the task and collected the data files every day.

### Results

**Descriptive statistics** Descriptive statistics for the average  $K$  values across the 31 sessions are shown in Table 2. Across all sessions, the average capacity was 2.83 ( $SD = 0.23$ ). Change in mean capacity over time is shown in Fig. 4a. A repeated measures ANOVA revealed a significant difference in capacity across sessions,  $F(18.76, 1388.38)^1 = 15.04, p < .001, \eta_p^2 = .169$ . Participants’ performance initially improved across sessions, then leveled off. The group-average increase in capacity over time is well-described by a two-term exponential model ( $SSE = .08, RMSE = .06, \text{adjusted } R^2 = .94$ ), described by the equation  $y = 2.776 \times e^{.003x} - 0.798 \times e^{-.26x}$ . To test the impression that individuals’ improvement slowed over time, we fit several growth curve models to the data using maximum likelihood estimation (fitmle.m) with Subject entered as a random factor. We coded time as days from the first session (Session 1 = 0). Model A included only a random intercept, Model B included a random intercept and a random linear

<sup>1</sup> Greenhouse–Geisser values reported when Mauchly’s test of sphericity is violated.

**Table 2** Descriptive statistics for Experiment 2

	<i>N</i>	Mean	<i>SD</i>	Minimum	Maximum	Kurtosis	Skewness
Day 1	79	2.15	0.85	0.40	4.03	-0.69	0.24
Day 2	79	2.36	0.86	0.07	3.97	-0.24	-0.32
Day 3	79	2.43	0.82	0.80	4.07	-0.62	-0.29
Day 4	<b>78</b>	2.51	0.85	0.40	4.10	-0.31	-0.31
Day 5	79	2.52	0.93	0.57	4.27	-0.55	-0.13
Day 6	79	2.74	0.92	0.53	4.60	-0.39	-0.20
Day 7	79	2.73	0.91	0.67	4.63	-0.88	-0.09
Day 8	79	2.66	0.87	1.03	4.70	-0.66	0.06
Day 9	79	2.81	0.92	0.50	5.07	-0.18	-0.19
Day 10	79	2.86	0.94	0.77	4.70	-0.84	0.01
Day 11	<b>78</b>	2.79	0.94	0.40	4.27	-0.51	-0.55*
Day 12	79	2.83	1.01	-0.10	4.80	-0.38	-0.37
Day 13	<b>78</b>	2.85	0.96	0.37	4.80	-0.57	-0.21
Day 14	79	3.01	0.95	0.93	5.03	-0.46	-0.11
Day 15	<b>78</b>	2.85	0.92	0.37	4.37	0.12	-0.73*
Day 16	79	2.91	0.92	0.23	4.90	-0.05	-0.35
Day 17	79	2.84	0.90	0.87	4.77	-0.51	-0.18
Day 18	79	2.93	1.02	0.53	4.73	-0.40	-0.23
Day 19	79	2.90	0.92	0.87	4.57	-0.69	-0.24
Day 20	79	2.94	0.92	0.47	4.93	-0.03	-0.32
Day 21	79	2.98	0.94	0.80	4.90	-0.08	-0.47
Day 22	79	2.99	0.98	0.83	4.90	-0.65	-0.23
Day 23	79	2.86	1.05	0.23	5.47	-0.17	-0.14
Day 24	<b>78</b>	3.00	0.98	0.97	4.77	-0.74	-0.26
Day 25	79	3.04	0.95	0.67	5.03	-0.41	-0.16
Day 26	79	3.01	0.93	0.43	5.07	-0.28	-0.34
Day 27	79	3.09	1.06	0.43	5.00	-0.51	-0.29
Day 28	79	3.04	0.97	0.33	4.83	-0.22	-0.48
Day 29	79	3.01	1.04	0.77	5.07	-0.38	-0.33
Day 30	79	3.02	1.05	0.33	5.00	-0.48	-0.29
Day 60	79	3.00	1.08	-0.13	5.40	0.29	-0.58*

Descriptive statistics are shown separately for each set size and for the average of the three set sizes. Kurtosis and skewness values are both centered around 0. Asterisks denote credible deviations from normality (Cramer, 1997). Lower *N* values for a few of the days are highlighted in bold.

effect of time, Model C added in a quadratic effect of time, and Model D added a cubic effect of time. As is shown in Table 3, the quadratic model provided the best fit to the data. Further testing revealed that both random slopes and intercepts were needed to best fit the data (see Table 4, comparing Models C1–C4). That is, participants started out with different baseline capacity values, and they improved at different rates. However, the covariance matrix for Model C revealed no systematic relationship between initial capacity (intercept) and either the linear effect of time,  $r = .21$ , 95% CI [-0.10, .49] or the quadratic effect of time,  $r = -.14$ , 95% CI [-0.48, .24]. This suggests that there was no meaningful relationship between a

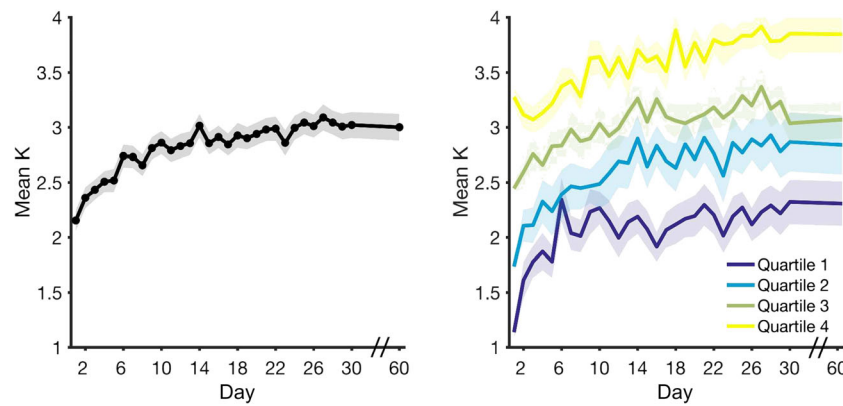
participant's initial capacity and that participant's rate of improvement. To visualize this point, we did a quartile split of Session 1 performance and then plotted the change for each of each group (Fig. 4).

**Within-session reliability** Within-session reliability was assessed using Cronbach's alpha and split-half correlations. Cronbach's alpha (using single-trial accuracy as items) yielded an average within-session reliability of  $\alpha = .76$  ( $SD = .04$ , min. = .65, max. = .83). Equivalently, split-half correlations on *K* scores calculated from even versus odd trials revealed an average Spearman–Brown-corrected reliability of  $r = .76$  ( $SD = .06$ , min. = .62, max. = .84). As in Experiment 1, using raw errors (Cronbach's alpha) versus bias-adjusted capacity measures (Cowan's *K*) did not affect the reliability estimates. Within-session reliability increased slightly over time (Fig. 5). Cronbach's alpha values were positively correlated with session number (1–31),  $r = .82$ ,  $p < .001$ , 95% CI [.66, .91], as were the split-half correlation values,  $r = .67$ ,  $p < .001$ , 95% CI [.42, .83].

**Between-session stability** We first assessed stability over time by computing correlation coefficients for all pairwise combinations of sessions (465 total combinations). Missing sessions were excluded from the correlations, meaning that some pairwise correlations included 78 participants instead of 79 (see Table 2). All sessions correlated with each other, mean  $r = .71$  ( $SD = .06$ , min. = .48, max. = .86, all  $p$  values < .001). A heat map of all pairwise correlations is shown in Fig. 6. Note that the most temporally distant sessions still correlated with each other. The correlation between Day 1 and Day 30 (28 intervening sessions) was  $r = .53$ ,  $p < .001$ , 95% CI [.35, .67]; the correlation between Day 30 and Day 60 (no intervening sessions) was  $r = .81$ ,  $p < .001$ , 95% CI [.72, .88]; the correlation between Day 1 and Day 60 was  $r = .59$ ,  $p < .001$ , 95% CI [.42, .71]. Finally, we observed that between-session stability increased over time, likely due to increased internal reliability across sessions. To compute change in reliability over time, we calculated the correlation coefficient for temporally adjacent sessions (e.g., the correlations of Session 1 and Session 2, of Session 2 and Session 3, etc.). The average adjacent-session correlation was  $r = .76$  ( $SD = .05$ , min. = .64, max. = .86), and the strength of adjacent-session correlations was positively correlated with session number,  $r = .68$ ,  $p < .001$ , indicating an increase in stability over time.

**Differences by testing location** We tested for systematic differences in performance, reliability, and stability for sessions completed at home versus in the lab. In total, 41 of the participants completed all of their sessions in their own home ("home group"), 27 participants completed all of their





**Fig. 4** Average capacity (*K*) across testing sessions. Shaded areas represent standard error of the mean. Note that the axis is spliced between Days 30 and 60, because no intervening data points were

collected during this time. (Left) Average changes in performance over time. (Right) Average changes in performance over time for each quartile of participants (quartile splits were performed on data from Session 1)

sessions in the lab (“lab group”), and 11 participants completed some sessions at home and some in the lab (“mixed group”).

Across all 31 sessions, the participants in the home group had an average capacity of 2.67 (*SD* = 1.01); those in the lab group had an average capacity of 3.01 (*SD* = 0.83); and those in the mixed group had an average capacity of 2.98 (*SD* = 1.04). On average, scores for sessions in the home group were slightly lower than scores for sessions in the lab group,  $t(2101) = -7.98$ ,  $p < .001$ , 95% CI [-.42, -.25]. Scores for sessions in the mixed group were higher than those for sessions in the home group,  $t(1606) = 5.0$ ,  $p < .001$ , 95% CI [.19, .43] but were not different from scores in the lab group,  $t(1175) = 0.44$ ,  $p = .67$ , 95% CI [-.09, .14]. Interestingly, however, a paired *t* test for the mixed group ( $n = 11$ ) revealed that the same participants performed slightly better in the lab ( $M = 3.08$ ) and slightly worse at home,  $M = 2.85$ ,  $t(10) = 3.15$ ,  $p = .01$ , 95% CI [.07, .39].

Cronbach’s alpha estimates of within-session reliability were slightly higher for sessions completed at home (mean  $\alpha = .76$ , *SD* = .05) than for sessions completed in the lab (mean  $\alpha = .69$ , *SD* = .08),  $t(60) = 3.75$ ,  $p < .001$ , 95% CI [.03, .10]. Likewise, Spearman–Brown-corrected correlation coefficients were higher for sessions completed at home (mean  $r = .79$ , *SD* = .07) than for those in the lab (mean  $r =$

.67, *SD* = .14),  $t(60) = 4.42$ ,  $p < .001$ , 95% CI [.07, .18]. However, these differences in reliability may have resulted from (1) unequal sample sizes between lab and home, (2) unequal average capacities between groups, or (3) unequal variabilities between groups. Once we equated sample sizes between groups and matched samples for average capacity, the differences in reliability were no longer stable: Across iterations of the matched samples, differences in Cronbach’s *alpha* ranged from  $p < .01$  to  $p > .5$ , and differences in split-half correlation significance ranged from  $p < .01$  to  $p > .25$ .

Next, we examined differences in stability for sessions completed at home rather than in the lab. On average, test–retest correlations were higher for home sessions (mean  $r = .72$ , *SD* = .08) than for lab sessions (mean  $r = .67$ , *SD* = .10),  $t(928) = 8.01$ ,  $p < .001$ , 95% CI [.04, .06]. Again, however, differences in the test–retest correlations were not reliable after matching sample size and average capacity; differences in correlation significance ranged from  $p = .01$  to .98.

**Discussion**

With extensive practice over multiple sessions, we observed improvement in overall change detection performance. This improvement was most pronounced over early sessions, after which mean performance stabilized for the remaining

**Table 3** Comparison of linear, quadratic, and cubic growth models, all with random intercepts and slopes where applicable

	Model A: Intercept Only	Model B: Linear	Model C: Quadratic	Model D: Cubic
Intercept	2.83***	2.60***	2.42***	2.29***
Linear slope		0.014***	.037***	.07**
Quadratic slope			-.0005***	-.002*
Cubic slope				$2 \times 10^{-5}$ n.s.
-2LL	4,366.2	4,084.8	3,914.7	4,231.6
BIC	4,389.6	4,131.6	3,992.7	4,348.6

\*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$

**Table 4** Comparison of fixed versus random slopes and intercept

	Model C1: Fixed Int. Fixed Slope	Model C2: Fixed Int. Random Slope	Model C3: Random Int. Fixed Slope	Model C4: Random Int. Random Slope
-2LL	6,672.3	4,627.7	4,009.1	3,914.7
BIC	6,703.5	4,682.3	4,048.1	3,992.7

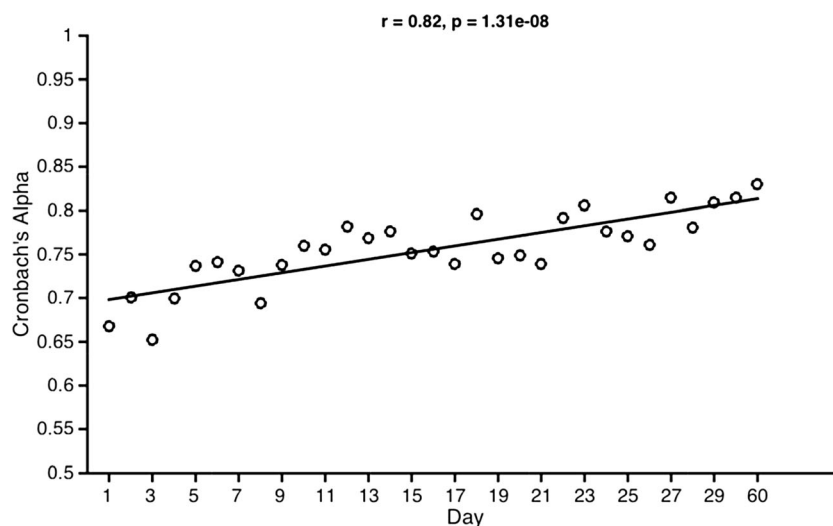
sessions. The internal reliability of the first session (Spearman–Brown-corrected  $r = .71$ , Cronbach's  $\alpha = .67$ ) was within the range predicted by the look-up table created in Experiment 1 for 80 participants and 120 trials (predicted range:  $r = .61$  to  $.87$  and  $\alpha = .58$  to  $.80$ , respectively). Both reliability and stability remained high over the span of 60 days. In fact, reliability and stability increased slightly across sessions. An important consideration for any cognitive measure is whether or not repeated exposure to the task will harm the reliability of the measure. For example, re-exposure to the same logic puzzles will drastically reduce the amount of time needed to solve the puzzles and inflate accuracy. Thus, for such tasks great care must be taken to generate novel test versions to be administered at different dates. Similarly, over-practice effects could lead to a sharp decrease in variability of performance (e.g., ceiling effects, floor effects), which would by definition lead to a decrease in reliability. Here, we demonstrated that although capacity estimates increase when participants are frequently exposed to a change detection task, the reliability of the measure is not compromised by either practice effects or ceiling effects.

We also examined whether reliability was harmed for participants who completed the change detection sessions in their own homes as compared to those in the lab. Although remote data collection sacrifices some degree of experimental control, the use of at-home tests is becoming more common with the ease of remote data collection through resources like

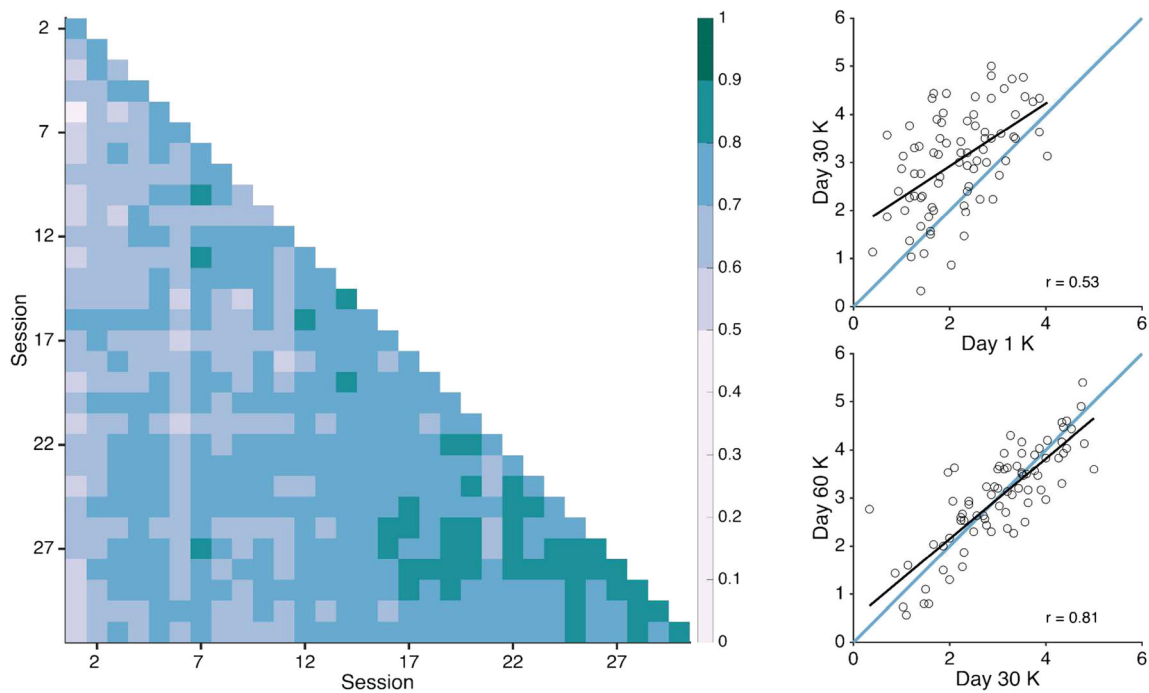
Amazon's Mechanical Turk (Mason & Suri, 2012). Reliability was not noticeably disrupted by noise arising from small differences in stimulus size between different testing environments. After controlling for number of participants and capacity, there was no longer a consistent difference in reliability or stability for sessions completed at home as compared to in the lab. However, the capacity estimates obtained in participants' homes were significantly lower than those obtained in the lab. Larger sample sizes will be needed to more fully investigate systematic differences in capacity and reliability between testing environments.

## General discussion

In Experiment 1, we developed a novel approach for estimating expected reliability in future experiments. We collected change detection data from a large number of participants and trials, and then we used an iterative down-sampling procedure to investigate the effect of sample size and trial number on reliability. Average reliability across iterations was fairly impervious to the number of participants. Instead, average reliability estimates across iterations relied more heavily on the number of trials per participant. On the other hand, the *variability* of reliability estimates across iterations was highly sensitive to the number of participants. For example, with only ten participants, the average reliability estimate for an experiment with 150 trials was high ( $\alpha = .75$ ) but the worst iteration (akin to the worst expected experiment out of 100) gave a poor reliability estimate ( $\alpha = .42$ ). On the other hand, the range between the best and worst reliability estimates decreased dramatically as the number of participants increased. With 40 participants, the minimum observed reliability for 150 trials was  $\alpha = .65$ .



**Fig. 5** Change in within-session reliabilities across sessions in Experiment 2. There was a significant positive relationship between session number (1:31) and internal reliability



**Fig. 6** Correlations between sessions. (Left) Correlations between all possible pairs of sessions. Colors represent the correlation coefficients of the capacity estimates from each possible pairwise combination of the 31 sessions. All correlation values were significant,  $p < .001$ .

In Experiment 2, we examined the reliability and stability of change detection capacity estimates across an unprecedented number of testing sessions. Participants completed 31 sessions of single-probe change detection. The first 30 sessions took place over 30 consecutive days, and the last session took place 30 days later (Day 60). Average internal reliability for the first session was in the range predicted by the look-up table in Experiment 1. Despite improvements in performance across sessions, between-subjects variability in  $K$  remained stable over time (average test–retest between all 31 sessions was  $r = .76$ ; the correlation for the two most distant sessions, Day 1 and Day 60, was  $r = .59$ ). Interestingly, both within-session reliability and between-session reliability increased across sessions. Rather than diminishing due to practice, reliability of WM capacity estimates increased across many sessions.

The present work has implications for planning studies with novel measures and for justifying the inclusion of existing measures into clinical batteries such as the Research Domain Criteria (RDoC) project (Cuthbert & Kozak, 2013; Rodebaugh et al., 2016). For basic research, an internal reliability of .7 is considered a sufficient “rule of thumb” for investigating correlational relationship between measures (Nunnally, 1978). Although this level of reliability (or even lower) will allow researchers to detect correlations, it is not sufficient to confidently assess the scores of individuals. For that, reliability in excess of .9 or even .95 is desirable (Nunnally, 1978). Here, we demonstrate how the number of

(Right) Illustration of the sessions that were most distant in time: Day 1 correlated with Day 30 (28 intervening sessions), and Day 30 correlated with Day 60 (no intervening sessions)

trials can alter the reliability of WM capacity estimates; with relatively few trials (~150, around 10 min of task time), change detection estimates are sufficiently reliable for correlation studies ( $\alpha \sim .8$ ), but many more trials are needed (~500) to boost reliability to the level needed to assess individuals ( $\alpha \sim .9$ ). Another important consideration for a diagnostic measure is its reliability across multiple testing sessions. Some tasks lose their diagnostic value once individuals have been exposed to them once or twice. Here we demonstrate that change detection estimates of WM capacity are stable, even when participants are well-practiced on the task (3,720 trials over 31 sessions).

One challenge in estimating the “true” reliability of a cognitive task is that reliability depends heavily on sample characteristics. As we have demonstrated, varying the sample size and number of trials can yield very different estimates of the reliability for a perfectly identical task. Other sample characteristics can likewise affect reliability; the most notable of these is sample homogeneity. The sample used here was a large sample of university students, with a fairly wide range in capacities (approximately 0.5–4 items). Samples using only a subset of this capacity range (e.g., clinical patient groups with very low capacity) will be less internally reliable because of the restricted range of the subpopulation. Indeed, in Experiment 1 we found that sampling iterations with poor reliability tended to have lower variability and a smaller range of scores. Thus, carefully recording sample size, mean, standard deviation, and internal reliability in all experiments will

be critical for assessing and improving the reliability of standardized tasks used for cognitive research. In the interest of replicability, open source code repositories (e.g., the Experiment Factory) have sought to make standardized versions of common cognitive tasks better-categorized, open, and easily available (Sochat et al., 2016). However, one potential weakness for task repositories is a lack of documentation about expected internal reliability. Standardization of tasks can be very useful, but it should not be over-applied. In particular, experiments with different goals should use different test lengths that best suit the goals of the experimental question. We feel that projects such as the Experiment Factory will certainly lead to more replicable science, and including estimates of reliability with task code could help to further this goal.

Finally, the results presented here have implications for researchers who are interested in differences between experimental conditions and not individual differences per se. Trial number and sample size will affect the degree of measurement error for each condition used within change detection experiments (e.g., set sizes, distractor presence, etc.). To detect significant differences between conditions and avoid false positives, it would be desirable to estimate the number of trials needed to ensure adequate internal reliability for each condition of interest within the experiment. Insufficient trial numbers or sample sizes can lead to intolerably low internal reliability, and could spoil an otherwise well-planned experiment.

The results of Experiments 1 and 2 revealed that change detection capacity estimates of visual WM capacity are both internally reliable and stable across many testing sessions. This finding is consistent with previous studies showing that other measures of WM capacity are reliable and stable, including complex span measures (Beckmann, Holling, & Kuhn, 2007; Foster et al., 2015; Klein & Fiss, 1999; Waters & Caplan, 1996) and the visuospatial *n*-back (Hockey & Geffen, 2004). The main analyses from Experiment 1 suggest concrete guidelines for designing studies that require reliable estimates of change detection capacity. When both sample size and trial numbers were high, the reliability of change detection was quite high ( $\alpha > .9$ ). However, studies with insufficient sample sizes or number of trials frequently had low internal reliability. Consistent with the notion that WM capacity is a stable trait of the individual, individual differences in capacity remained stable over many sessions in Experiment 2 despite practice-related performance increases.

Both the effects of trial number and sample size are important to consider, and researchers should be cautious about generalizing expected reliability across vastly different sample sizes. For example, in a recent article by Foster and colleagues (2015), the authors found that cutting the number of complex span trials by two-thirds had only a modest effect on the strength of the correlation between WM capacity and fluid intelligence. Critically, however, the authors used around

500 participants, and such a large sample size will act as a buffer against increases in measurement error (i.e., fewer trials per participant). Readers wishing to conduct a new study with a smaller sample size (e.g., 50 participants) would be ill-advised to dramatically cut trial numbers on the basis of this finding alone; as we demonstrated in Experiment 1, cutting trial numbers leads to greater volatility of reliability values for small sample sizes relative to large ones. Given the current concerns about power and replicability in psychological research (Open Science Collaboration, 2015), we suggest that rigorous estimations of task reliability, considering both participant and trial numbers, will be useful for planning both new studies and replication efforts.

#### Acknowledgements

**Contributions** Z.X. and E.V. designed the experiments; Z.X. and X.F. collected data. K.A. performed the analyses and drafted the manuscript, and K.A., Z.X., and E.V. revised the manuscript.

**Author note** Research was supported by the Project of Humanities and Social Sciences, Ministry of Education, China (15YJA190008), the Fundamental Research Funds for the Central Universities (SWU1309117), NIH Grant 2R01 MH087214-06A1, and Office of Naval Research Grant N00014-12-1-0972. Datasets for all experiments are available online on Open Science Framework at <https://osf.io/g7xf/>.

#### Compliance with ethical standards

**Conflicts of interest** None.

#### References

- Beckmann, B., Holling, H., & Kuhn, J.-T. (2007). Reliability of verbal–numerical working memory tasks. *Personality and Individual Differences, 43*, 703–714. doi:10.1016/j.paid.2007.01.011
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436. doi:10.1163/156856897X00357
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904–1920*(3), 296–322. doi:10.1111/j.2044-8295.1910.tb00207.x
- Buschman, T. J., Siegel, M., Roy, J. E., & Miller, E. K. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences, 108*, 11252–11255. doi:10.1073/pnas.1104666108
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences 24*, 87–114–185. doi:10.1017/S0140525X01003922
- Cowan, N., Fristoe, N. M., Elliott, E. M., Brunner, R. P., & Sauls, J. S. (2006). Scope of attention, control of attention, and intelligence in children and adults. *Memory & Cognition, 34*, 1754–1768. doi:10.3758/BF03195936
- Cramer, D. (1997). *Basic statistics for social research: Step-by-step calculations and computer techniques using Minitab*. London: Routledge.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi:10.1007/BF02310555



- Cuthbert, B. N., & Kozak, M. J. (2013). Constructing constructs for psychopathology: The NIMH research domain criteria. *Journal of Abnormal Psychology, 122*, 928–937. doi:10.1037/a0034028
- Elmore, L. C., Magnotti, J. F., Katz, J. S., & Wright, A. A. (2012). Change detection by rhesus monkeys (*Macaca mulatta*) and pigeons (*Columba livia*). *Journal of Comparative Psychology, 126*, 203–212. doi:10.1037/a0026356
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*, 309–331. doi: 10.1037/0096-3445.128.3.309
- Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition, 43*, 226–236. doi:10.3758/s13421-014-0461-7
- 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. doi:10.3758/17.5.673
- Fukuda, K., Woodman, G. F., & Vogel, E. K. (2015). Individual differences in visual working memory capacity: Contributions of attentional control to storage. In P. Joliceur, C. Lefebvre, & J. Martinez-Trujillo (Eds.), *Mechanisms of sensory working memory: Attention and performance XXV* (pp. 105–119). San Diego: Academic Press Elsevier.
- Gibson, B., Wasserman, E., & Luck, S. J. (2011). Qualitative similarities in the visual short-term memory of pigeons and people. *Psychonomic Bulletin & Review, 18*, 979–984. doi:10.3758/s13423-011-0132-7
- Gold, J. M., Wilk, C. M., McMahon, R. P., Buchanan, R. W., & Luck, S. J. (2003). Working memory for visual features and conjunctions in schizophrenia. *Journal of Abnormal Psychology, 112*, 61–71. doi: 10.1037/0021-843X.112.1.61
- Hockey, A., & Geffen, G. (2004). The concurrent validity and test-retest reliability of a visuospatial working memory task. *Intelligence, 32*, 591–605. doi: 10.1016/j.intell.2004.07.009
- Johnson, M. K., McMahon, R. P., Robinson, B. M., Harvey, A. N., Hahn, B., Leonard, C. J., & Gold, J. M. (2013). The relationship between working memory capacity and broad measures of cognitive ability in healthy adults and people with schizophrenia. *Neuropsychology, 27*, 220–229. doi:10.1037/a0032060
- Klein, K., & Fiss, W. H. (1999). The reliability and stability of the Turner and Engle working memory task. *Behavior Research Methods, Instruments, & Computers, 31*, 429–432. doi:10.3758/BF03200722
- Lee, E.-Y., Cowan, N., Vogel, E. K., Rolan, T., Valle-Inclan, F., & Hackley, S. A. (2010). Visual working memory deficits in patients with Parkinson's disease are due to both reduced storage capacity and impaired ability to filter out irrelevant information. *Brain, 133*, 2677–2689. doi:10.1093/brain/awq197
- Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience & Biobehavioral Reviews, 62*, 100–108. doi:10.1016/j.neubiorev.2016.01.003
- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*, 1–23. doi:10.3758/s13428-011-0124-6
- Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49*, 270–291. doi:10.1037/a0028228
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. doi:10.1126/science.aac4716
- Pailian, H., & Halberda, J. (2015). The reliability and internal consistency of one-shot and flicker change detection for measuring individual differences in visual working memory capacity. *Memory & Cognition, 43*, 397–420. doi:10.3758/s13421-014-0492-0
- Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics, 44*, 369–378. doi:10.3758/BF03210419
- Reinhart, R. M. G., Heitz, R. P., Purcell, B. A., Weigand, P. K., Schall, J. D., & Woodman, G. F. (2012). Homologous mechanisms of visuospatial working memory maintenance in macaque and human: properties and sources. *Journal of Neuroscience, 32*, 7711–7722. doi:10.1523/JNEUROSCI.0215-12.2012
- Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., & Lenze, E. J. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. *Journal of Abnormal Psychology, 125*, 840–851. doi:10.1037/abn0000184
- Rouder, J. N. (n.d.). Applications and source code. Retrieved June 22, 2016, from <http://pcl.missouri.edu/apps>
- Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences, 105*, 5975–5979. doi:10.1073/pnas.0711295105
- 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. doi:10.3758/s13423-011-0055-3
- Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin, 138*, 628–654. doi:10.1037/a0027473
- Sochat, V. V., Eisenberg, I. W., Enkavi, A. Z., Li, J., Bissett, P. G., & Poldrack, R. A. (2016). The experiment factory: Standardizing behavioral experiments. *Frontiers in Psychology, 7*, 610. doi:10.3389/fpsyg.2016.00610
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904–1920(3)*, 271–295. doi:10.1111/j.2044-8295.1910.tb00206.x
- Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature, 428*, 751–754. doi:10.1038/nature02466
- 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. doi:10.1016/j.cogpsych.2014.01.003
- Van Snellenberg, J. X., Conway, A. R. A., Spicer, J., Read, C., & Smith, E. E. (2014). Capacity estimates in working memory: Reliability and interrelationships among tasks. *Cognitive, Affective, & Behavioral Neuroscience, 14*, 106–116. doi:10.3758/s13415-013-0235-x
- Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature, 428*, 748–751. doi:10.1038/nature02447
- Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *Quarterly Journal of Experimental Psychology, 49A*, 51–75. doi: 10.1080/713755607
- Wood, G., Hartley, G., Furley, P. A., & Wilson, M. R. (2016). Working memory capacity, visual attention and hazard perception in driving. *Journal of Applied Research in Memory and Cognition, 5*, 454–462. doi:10.1016/j.jarmac.2016.04.009