Change localization: A highly reliable and sensitive measure of capacity in visual working memory

Chong Zhao[1,2] · Edward Vogel[1,2] · Edward Awh[1,2]

Abstract

The change detection paradigm has been a widely used approach for measuring capacity in visual working memory (WM). In this task, subjects see an array of visual items, followed by a short blank delay and a single test item. Their task is to indicate whether that test item changed relative to the item in the sample array. This task provides reliable measurements of WM capacity that exhibit robust correlations with many outcome variables of interest. Here, we offer a new variant of this task that we call "change localization." This task is closely modeled after the change detection task described above, except that the test array contains the same number of items as the sample array, and one item has always changed in each trial. The subject's task is to select the changed item in the test array. Using both color and shape stimuli, scores in the change localization task were highly correlated with those in the change detection task, suggesting that change localization taps into the same variance in WM ability. Moreover, the change localization task was far more reliable than change detection, such that only half the number of trials were required to achieve robust reliability. To further validate the approach, we replicated known effects from the literature, demonstrating that they could be detected with far fewer trials than with change detection. Thus, change localization provides a highly reliable and sensitive approach for measuring visual working memory capacity.

Keywords Visual working memory · Change localization paradigm · Reliability

Introduction

Individual differences in visual working memory (WM) capacity predict a wide range of cognitive abilities, such as verbal memory encoding (Miller et al., 2019), mathematical reasoning (Raghubar et al., 2010), and fluid intelligence (Unsworth et al., 2014). Thus, there is clear motivation to develop robust and reliable behavioral measures of this cognitive ability. One influential approach is the so-called "change detection" procedure used by Luck and Vogel (1997) to measure capacity limits in visual WM. In this task, subjects are shown an array of colored squares, and then decide whether any item has changed in a test display that is presented after a brief blank delay period. Change detection

has been shown to provide highly reliable measurements of individual differences in WM capacity (Xu et al., 2018), and has helped to reveal the links between WM ability and variations in attentional control, fluid intelligence, and long-term memory access (Unsworth et al., 2014). Nevertheless, we argue here that a small modification of this procedure – from change detection to change *localization* – may provide substantial improvements in reliability and sensitivity. In turn, this may allow more efficient and higher quality tracking of WM capacity across different experimental conditions.

Our examination of change localization was motivated by a past study (Pailian & Halberda, 2015) that had found evidence of higher reliability in a task that required observers to localize the changed item. Here, we collected a much larger sample size that enabled a comprehensive analysis of how reliability in these tasks varies as a function of the number of subjects and trials. In our change localization task, one item changed in every test display, and observers localized that item. Because there is no threshold for detecting changes, this measure removes concerns about response bias, a factor that can reduce the precision of capacity estimates (Williams et al., 2022). In addition, because chance performance is lower for change localization (1/6) than for change detection (1/2), change localization provides a greater

✉ Chong Zhao
chongzhao@uchicago.edu

[1] Department of Psychology, University of Chicago, Chicago, Illinois 60637, USA

[2] Institute for Mind and Biology, University of Chicago, Chicago, Illinois 60637, USA

dynamic range for performance; this provides a better testbed for examining individual differences in working memory capacity. Finally, in addition to characterizing the reliability of the tasks, we carried out a direct comparison of their relative sensitivity to known empirical patterns.

To anticipate our findings, we observed a strong correlation between change localization and change detection scores, suggesting that change localization taps into the same capacity construct. Second, we used a downsampling approach to determine the number of trials and subjects needed to achieve high reliability (Xu et al., 2018). This analysis revealed a substantial advantage in reliability for change localization compared to change detection, such that only half as many trials of change localization were required to achieve the same reliability scores. Finally, change localization required far fewer trials to detect known empirical patterns from the change detection literature, suggesting that this new task may provide a more efficient approach for a wide range of questions.

Experiment 1

Method

Participants

One hundred participants were recruited at the rate of $9.50 per hour from Prolific, an online platform for participant recruitment. All participants were 18–35 years old, currently lived in the USA, had normal or corrected-to-normal vision, and had no ongoing psychological or neurological disorders.

Stimuli

In both of our visual working memory tasks, all stimuli were colored squares generated in Javascript using the jsPsych canvas-keyboard interface. The colored squares were all in 40 x 40 pixel size on a 400 x 400 pixel canvas page. Color squares could appear anywhere within a circular area of the monitor within 30–200 pixels from the center of the canvas screen. Each square could appear in one of the nine distinct colors with no repetitions within any 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 at a small black plus (30 px in Arial) at the center of the screen throughout the trial.

Procedures

Change detection In each trial, six colored squares appeared on the screen simultaneously for 250 ms, followed by a 1,000-ms blank retention interval. Then, one colored square

appeared and participants indicated whether the color of this square had changed compared to the square previously presented in that position. Changes occurred with a probability of 0.5. Each participant completed 240 trials of the change detection task in total (see Fig. 1 Upper).

Change localization In each trial, six colored squares appeared simultaneously for 250 ms, followed by a 1,000-ms blank retention interval. Then, the six squares were presented again in the same locations, with one of the six colors changed to a color that had not been presented in the trial. Each square was marked with a digit (from 1 to 6), and participants pressed the corresponding key to indicate the item that had changed color. The spatial position of the six numbers was randomized across trials. Each participant completed 240 trials of the change localization task in total (see Fig. 1 Lower).
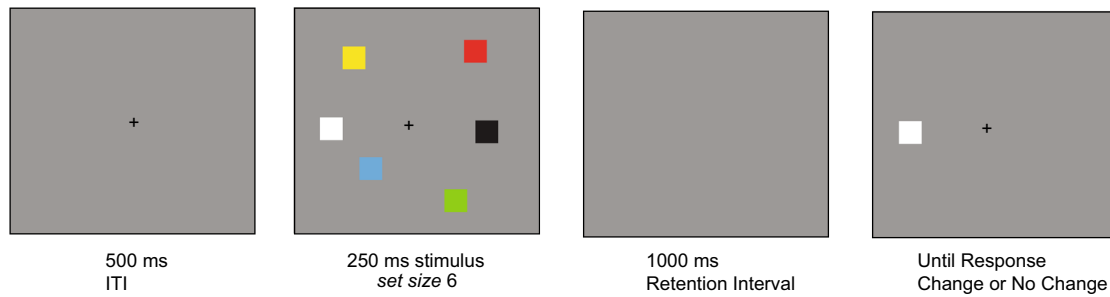
Results

The mean K capacity estimates using the change detection paradigm was 2.45, similar to the K scores in previous literatures (Balaban et al., 2019). In our change localization paradigm, the mean accuracy was 58.6%, much higher than the statistical chance at 16.67%. To test the reliability of each paradigm, we performed an odd-even split-half analysis. This analysis revealed that the change detection paradigm had a high internal reliability ($r(99) = 0.87$, $p = 2.19 \times 10^{-32}$). Similarly, the change localization paradigm also exhibited a high internal reliability ($r(99) = 0.93$, $p = 1.38 \times 10^{-44}$). We then examined if change localization accounted for the same variance in visual working memory as change detection. A strong correlation between the two tasks suggested that they were tapping into a common memory ability ($r(99) = 0.82$, $p = 1.55 \times 10^{-25}$, see Fig. 2).
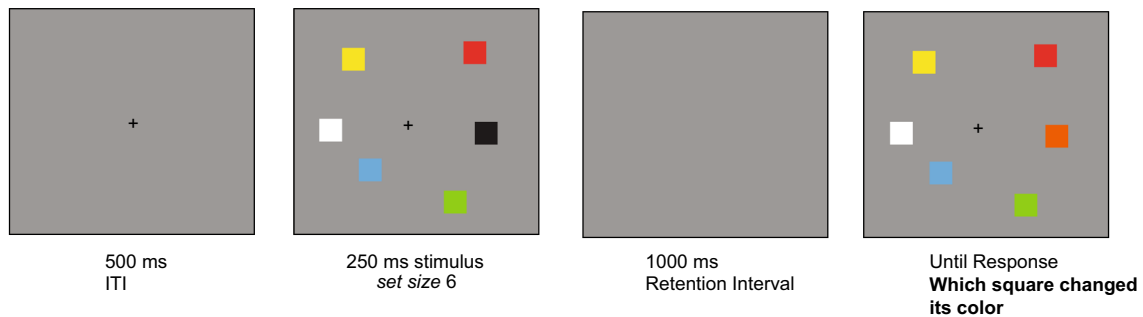
Change localization paradigm was statistically more reliable than change detection paradigm

To compare the reliability of change detection and change localization, we performed iterative down-sampling on the split-half correlations of the two paradigms. First, we down-sampled trials and subjects from the whole dataset, starting from the smallest number and increasing until the whole sample was used. Then, we could calculate, for each unique combination of subject and trial numbers, the reliability metric within these smaller samples. This analysis reveals the lowest possible numbers of subjects and trials that are needed to achieve a given level of reliability. We ran 100 iterations for each combination of trials and subject numbers. We randomly sampled the number of subjects from five to 100 in steps of five, and the number of trials from five to 200 in steps of five. One obvious result observed from the down-sampling figures was that change

Task A: Simultaneous Change Detection Paradigm (240 trials)

500 ms ITI	250 ms stimulus *set size* 6	1000 ms Retention Interval	Until Response Change or No Change

Task B: Simultaneous Change Localization Paradigm (240 trials)

|			
500 ms ITI	250 ms stimulus *set size* 6	1000 ms Retention Interval	Until Response **Which square changed its color**

Fig. 1 The change detection paradigm and change localization paradigms used in Exp 1. (Upper) A sample of set size 6 simultaneous change detection paradigm. Six colored squares appeared on the screen simultaneously. At the end of the trial, participants were cued with a square at one of the six original locations, with 50% exhibited a color change and 50% without change. (Lower) A sample of the *change localization* paradigm.

Six colored squares appeared on the screen simultaneously. At the end of the trial, participants were cued with six squares at the original locations as the study phase. One of the squares would change its color, and participants need to report which one changed its color by pressing the keyboard buttons. Each square would have a number on it during the test phase for participants to press the corresponding button

localization required a smaller number of trials and subjects to reach statistical reliability, as illustrated by the broader swath of green in the change localization compared to the change detection figure (see Fig. 3a and b). To better quantify the advantage of change localization on reliability, we calculated the minimum number of trials required to reach a Spearman-Brown correlation of 0.8 in each paradigm. With more than 20 subjects, change localization required only half the number of trials that were required with change detection (see Fig. 3c).

Discussion

In Experiment 1, we replicated previous research in showing that the change detection task was highly reliable (Xu et al., 2018), but we found substantially higher reliability for the change localization task, such that only half the number of trials were needed to achieve the same reliability. In Experiment 2, we sought to replicate these findings with shape rather than color memoranda.
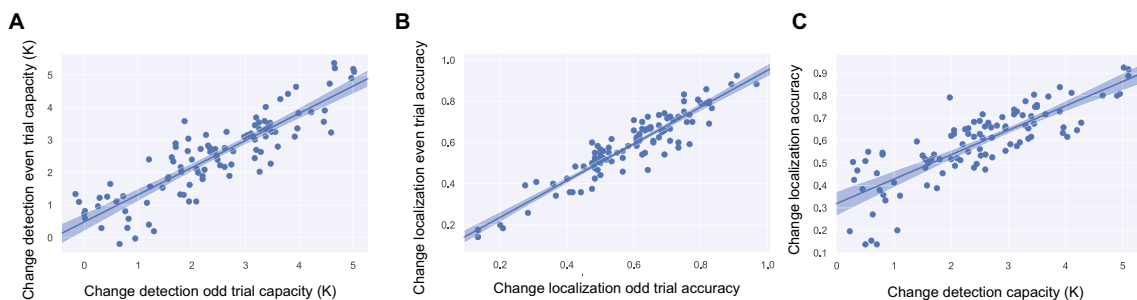
Fig. 2 Change detection and change localization paradigms both had high reliability. **a** An odd-even split-half analysis revealed that the change detection paradigm had a high internal reliability, with a Pearson correlation of 0.87. **b** An odd-even split-half analysis suggested that the change localization paradigm also had a high internal reliability, with a Pearson correlation of 0.93. **c** Change localization exhibited the same individual differences measured by change detection paradigm, with a Pearson correlation of 0.82

A Change detection
downsampling

B Change localization
downsampling

C Comparison
of reliability
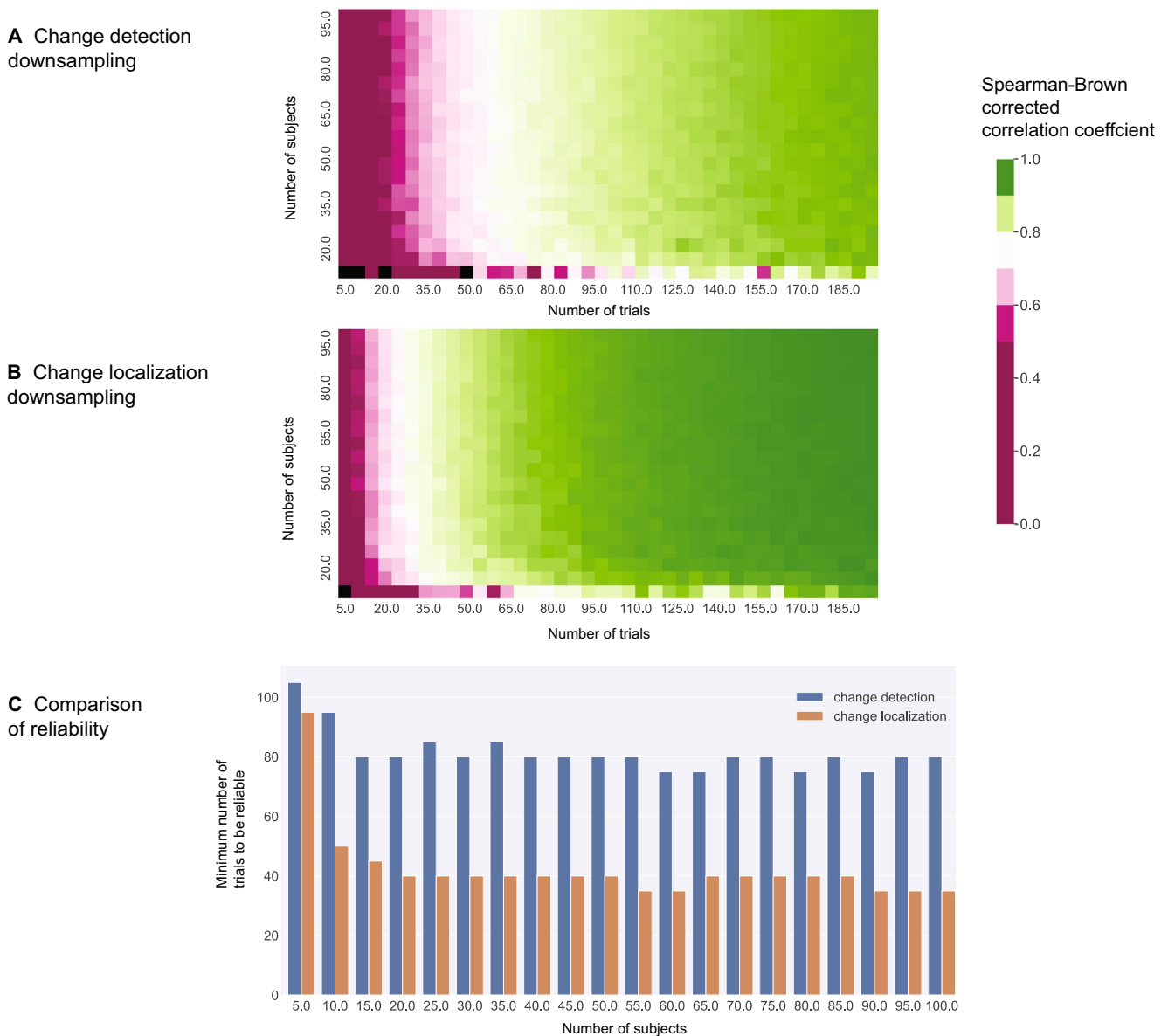
Fig. 3 Change localization was statistically more reliable than change detection paradigm. **a** Iterative downsampling of split-half correlation using change detection data. **b** Iterative downsampling of split-half correlation using change localization data. **c** The minimum number of trials that reached reliability goal (split-half corrected correlation of 0.8) in change detection and change localization paradigm. Across all sample sizes, change localization outperformed change detection by requiring far less trials to be similarly reliable

Experiment 2

Method

Participants

One hundred participants were recruited at the rate of $9.50 per hour from Prolific, an online platform for participant recruitment. All participants were 18–35 years old, currently lived in the USA, had normal or corrected-to-normal vision, and had no ongoing psychological or neurological disorders.

Stimuli

In both of our visual working memory tasks, we created eight different color shapes with the same gray background (circle, x, triangle, vertical bar, horizontal bar, square, crescent and cog). An example of all shapes in black is shown in Online Supplementary Material (OSM) Fig. 1. A critical design goal for these stimuli was to achieve the highest discriminability for all possible pairs of objects in the array. Excellent discriminability for each possible pair of shapes should minimize *comparison errors* in which the observer had stored the

changed item, but did not have a precise enough memory to detect the change. This is important because past work has shown that when comparison errors are prevalent, performance in such tasks begins to track the precision of memory rather than the number of items that can be retained (Awh et al., 2007; Fukuda et al., 2010).

Colored shapes could appear anywhere within a circular area of the monitor within 30–200 pixels from the center of the canvas screen. Each shape was randomly placed in one of the six sextants such that no two shapes were placed within the same 60° on the screen. Furthermore, the minimum distance between two neighboring shapes would not exceed 36.7 pixels. Each shape could appear in one of the nine distinct colors, and the color was the same for all stimuli within any trial (RGB values: red = 175 0 0; green = 0 185 0; blue = 0 0 255; magenta = 255 0 255; yellow = 255 255 0; cyan = 0 255 255; orange = 255 155 0; white = 255 255 255; black = 0 0 0). Participants were instructed to fixate at a small black plus (30 px in Arial) at the center of the screen throughout the trial.

Procedures

We administered a change detection task with set size 6. In each trial, six colored shapes appeared on the screen simultaneously for 250 ms, followed by a 1,000-ms blank retention interval. Then, one color shape appeared at the position of one of the prior shapes, and the participant indicated whether the shape had changed. Changes occurred with a probability of 0.5. Each participant completed 240 trials of the change detection task in total (see Fig. 4 Upper).

We also administered the change localization task with set size 6. In each trial, six colored shapes appeared simultaneously for 250 ms, followed by 1,000 ms of a blank retention interval. Then, the six shapes reappeared in the same positions, with one of the shapes having changed on every trial. The participant indicated which one of the six colored shapes had changed by typing the number of a digit that appeared at that shape's location. Each participant completed 240 trials of the change localization task in total (see Fig. 4 Lower).

Result

Change localization measured the same individual differences as revealed by change detection

To test the reliability of each paradigm, we performed an odd-even split-half analysis by dividing all trials into halves to calculate the reliability of each metric. The split-half analysis

Task A: Simultaneous Change Detection Paradigm (240 trials)
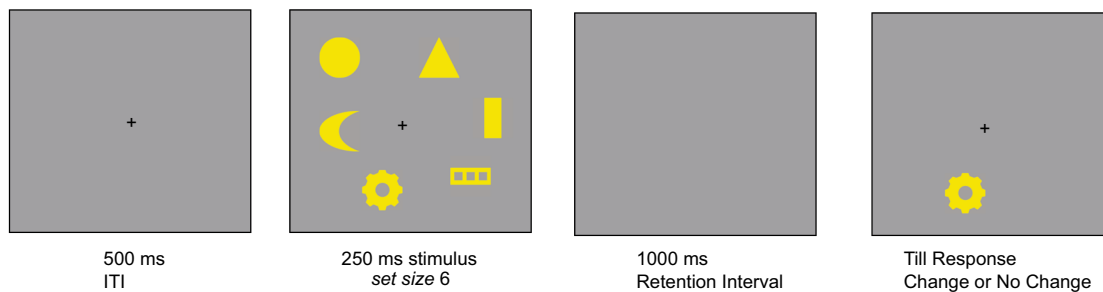
Task B: Simultaneous Change Localization Paradigm (240 trials)
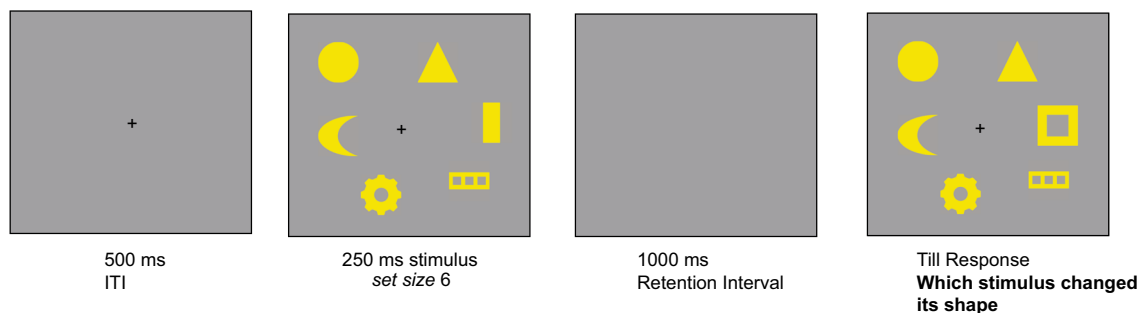
Fig. 4 The change detection paradigm and change localization paradigm used in Exp 2. (Upper) A sample of set size 6 simultaneous change detection paradigm. Six colored shapes appeared on the screen simultaneously. At the end of the trial, participants were cued with a shape at one of the six original locations, with 50% exhibited a shape change and 50% without change. (Lower) A sample of the *change localization* paradigm.

Six colored shapes appeared on the screen simultaneously. At the end of the trial, participants were cued with six shapes at the original locations as the study phase. One of the stimuli would change its shape. and participants need to report which one changed its shape by pressing the keyboard buttons. Each shape would have a number on it during the test phase for participants to press the corresponding button

revealed that change detection paradigm had a high internal reliability(r(99) = 0.85, p = 2.10 x 10^{-28}). Similarly, change localization also had high internal reliability (r(99) = 0.91, p = 1.63 x 10^{-38}). Given the high reliability of both paradigms, we then examined if change localization accounted for the same variance in visual working memory as change detection. A correlational analysis between two paradigms suggested that they were measuring strongly overlapping variance (r(99) = 0.65, p = 2.33 x 10^{-13}, see Fig. 5).

Change localization was statistically more reliable than change detection

To further examine if change localization was more statistically reliable, we performed iterative down-sampling on the split-half correlations of the two paradigms just like in Experiment 1. We first down-sampled trials and subjects from the whole dataset. Then, we could calculate, for each unique combination of subject and trial numbers, the reliability metric within these smaller samples. This analysis revealed the lowest possible numbers of subjects and trials needed to achieve a given level of reliability. In our experiment, we adopted 100 iterations of runs in the down-sampling procedure, meaning that each number of trials and subjects were randomly sampled for 100 times from our dataset. We randomly sampled the number of subjects from five to 100 in steps of five, and the number of trials from five to 200 in steps of five. One obvious result observed from the down-sampling figures was that change localization required far fewer trials and subjects to reach the same statistical reliability as change detection. This is illustrated by the broader swath of green points that indicate a reliability of at least 0.8. Figure 6a and b illustrates this with a much broader region of green points (indicating a reliability of at least 0.8) for change localization compared to change detection. To better quantify the advantage of change localization on reliability, we calculated the minimum number of trials required to reach a Spearman-Brown correlation of 0.8 in each paradigm. The result revealed that change localization always needed less trials to be statistically reliable than change detection,

and with more than ten subjects, it required only half (or fewer) trials as required by change detection (see Fig. 6c and Table 1).

Discussion

Experiment 2 showed that change detection task had high internal reliability when the feature being tested was shape instead of color. Furthermore, we found that that change localization task using shape was also statistically reliable with 240 trials and shared highly similar variance in measuring visual working memory performance as the change detection paradigm. Lastly, we found that change localization paradigm outperformed change detection paradigm in reliability in that it always required less than half the number of trials to be reliable using the same number of participants.

In our first two experiments, we showed that the change localization paradigm was highly reliable, and tapped into similar variance in performance as the change detection task. Both experiments utilized a relatively high set size so that individual differences were easier to detect. In our next experiment, we examined whether the change localization paradigm remained reliable with lower set sizes of only four or five items, and we examined its sensitivity to detecting the difference in performance between these two conditions.

Experiment 3

Method

Participants

Fifty-eight participants were recruited at the rate of $9.50 per hour from Prolific, an online platform for participant recruitment. All participants were 18–35 years old, currently lived in the USA, had normal or corrected-to-normal vision, and had no ongoing psychological or neurological disorders.
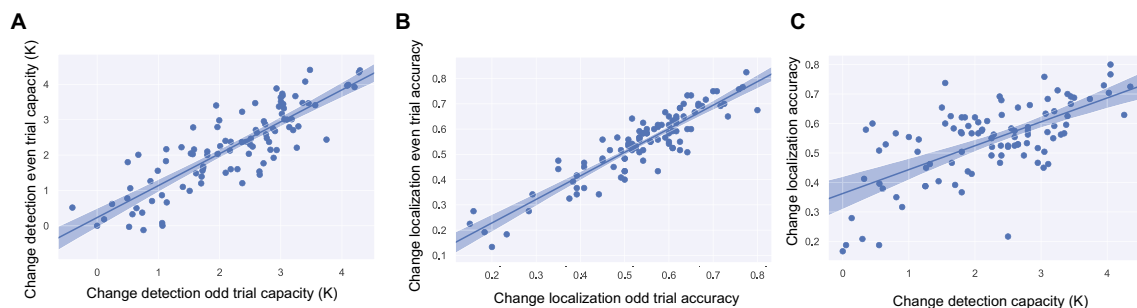
Fig. 5 Change detection and change localization paradigms both had high reliability. **a** An odd-even split-half analysis revealed that the change detection paradigm had a high internal reliability, with a Pearson correlation of 0.85. **b** An odd-even split-half analysis suggested that the change localization paradigm also had a high internal reliability, with a Pearson correlation of 0.91. **c** Change localization exhibited similar individual differences measured by change detection paradigm, with a Pearson correlation of 0.65

Table 1 Minimum trials required to achieve reliability with 20 subjects

| Minimum trials required to achieve reliability with 20 subjects	Change detection	Change localization
Exp. 1: color working memory task (ss6)	80	40
Exp. 2: shape working memory task (ss6)	135	60

Stimuli

The stimuli used were the same as in Experiment 1.

## Procedures

In measuring visual working memory capacity, we administered a color version of change detection task similar to that in Experiment 1. Each participant was asked to determine if the color of the probed item changed with four or five items during encoding, different from the set size 6 used in Experiment 1. Meanwhile, each participant also finished color change localization task with four or five test items on screen instead of set size 6 (other procedures as in Exp. 1). Overall, each participant finished 140 trials each of change localization with set
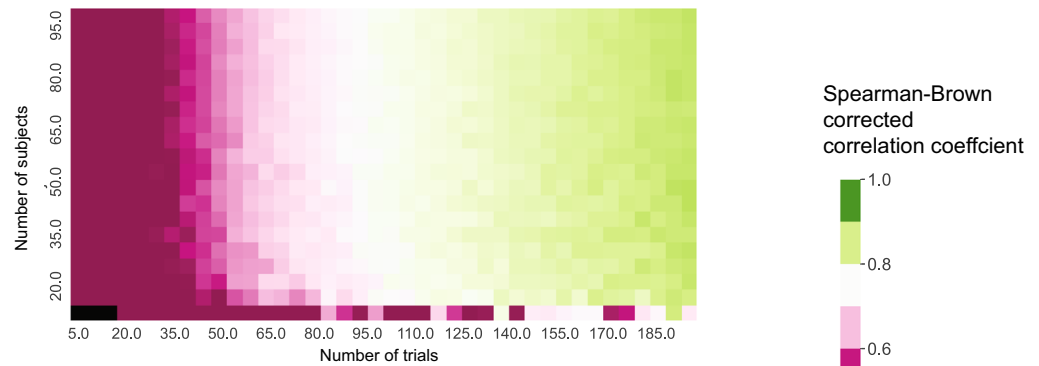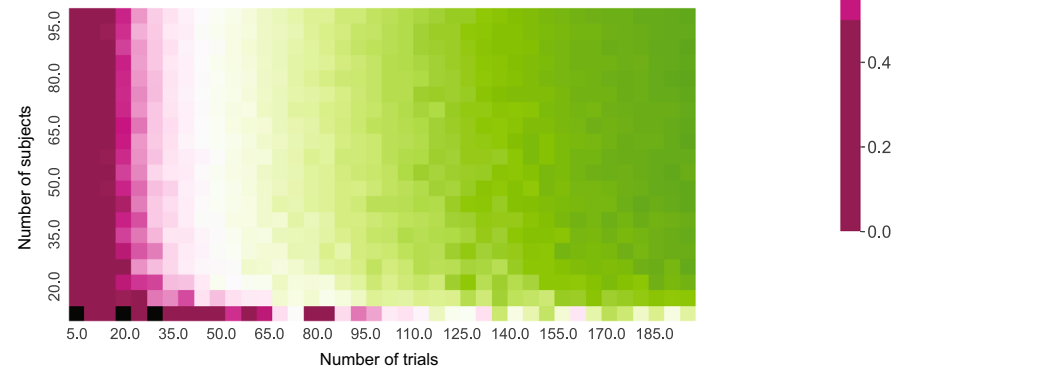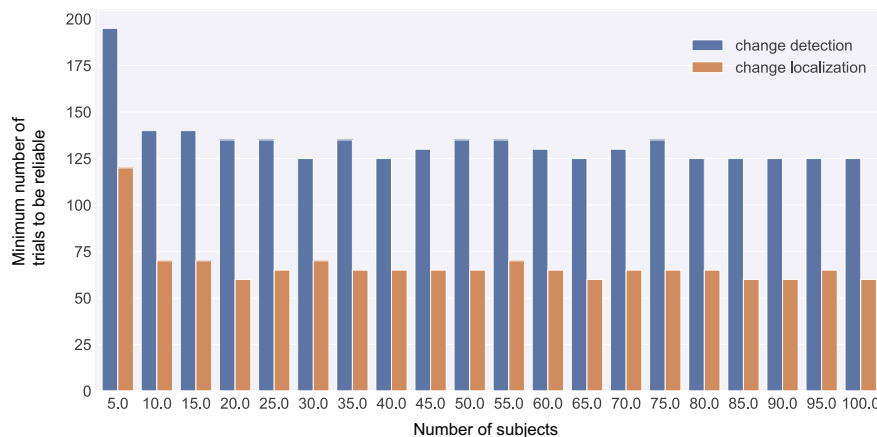

Fig. 6 Change localization was statistically more reliable than change detection paradigm. **a** Iterative downsampling of split-half correlation using change detection data. **b** Iterative downsampling of split-half correlation using change localization data. **c** The minimum number of trials that reached reliability goal (split-half corrected correlation of 0.8) in change detection and change localization paradigm. Across all sample sizes, change localization outperformed change detection by requiring far less trials to be similarly reliable

size 4 and 5, and 140 trials each of change detection with set size 4 and 5, resulting in 560 trials in total.

Result

Change localization measured similar individual differences as revealed by change detection paradigm across two smaller set sizes

To test the reliability of each paradigm, we performed an odd-even split-half analysis by dividing all trials into halves to calculate the reliability of each metric. The split-half analysis revealed that the change detection paradigm had a high internal reliability with both set size 4 ($r(56) = 0.87$, $p = 5.29$ x 10^{-19}) and set size 5 ($r(56) = 0.84$, $p = 3.09$ x 10^{-16}). Similarly, the change localization paradigm was also shown to have a high internal reliability with both set size 4 ($r(56) = 0.90$, $p = 1.76$ x 10^{-21}) and set size 5 ($r(56) = 0.90$, $p = 2.15$ x 10^{-22}). Given the high reliability of both paradigms, we then examined if change localization accounted for the same variance in visual working memory as change detection. A correlational analysis between two paradigms suggested that they were measuring similar variance under set size 4 ($r(56) = 0.71$, $p = 2.91$ x 10^{-10}) and set size 5 conditions ($r(56) = 0.78$, $p = 4.97$ x 10^{-13}, see Fig. 7a and b).

Change localization was more sensitive in capturing set size effects than change detection

To further examine if change localization was more sensitive in capturing set size accuracy differences, we performed iterative down-sampling on the t-statistics of the two paradigms just like in Experiment 1. We first down-sampled trials and subjects from the whole dataset. Then, we could calculate, for each unique combination of subject and trial numbers, the t statistics and resulting p-value between accuracies for set size 4 trials and set size 5 trials within these smaller samples. This analysis revealed the lowest possible numbers of subjects and trials needed to achieve a given level of significance value. In our experiment, we adopted 100 iterations of runs in the down-sampling procedure, meaning that each number of trials and subjects were randomly sampled for 100 times from our dataset. We randomly sampled the number of subjects from five to 50 in steps of 5, and the number of trials from five to 100 in steps of 5. One obvious result observed from the down-sampling figures was that change localization required far fewer trials and subjects to reach the same statistical sensitivity as change detection. This, as illustrated by the broader swath of blue points that indicate a p-value of at most 0.05. Figure 7b and c illustrates this with a much broader region of blue points (indicating a p-value of at most 0.05) for change localization compared to change detection.

Determining a formula for calculating K using change localization data

Past work has often quantified working memory capacity in terms of K, or the number of items stored on average in each trial. Here, our goal was to determine the appropriate formula to estimate K using change localization data. Here, it is critical to highlight a key assumption of this measurement approach. Specifically, the proposed formula presumes that the *limiting factor* for accurate performance is whether the changed item was *stored*, not whether the *fidelity* of stored memories was sufficient to detect small changes. The same assumption underlies each of the K formulas that have been offered in the extant literature (Cowan, 2001; Luck & Vogel, 1997; Pashler, 1988). Indeed, past work has shown that when comparison errors are prevalent, change detection performance may tap into a qualitatively distinct aspect of memory ability (Awh et al., 2007). For example, studies of individual differences have shown that separate latent variables emerge for detecting large and small changes (Fukuda et al., 2010; Unsworth et al., 2014), consistent with the hypothesis that these measures tap into the number and fidelity of the stored representations, respectively. Thus, we minimized comparison errors by choosing stimuli that maximize the discriminability of every possible pair of colors or shapes.

Assuming that a participant had a visual working memory capacity of K and the encoding array had N items in total, the probabilities for the participant to successfully maintain the test item in their working memory or fail to maintain the item were:

$$P(success) = \frac{K}{N}$$
$$P(fail) = \frac{(N-K)}{N}$$

respectively. We further assumed that if the participant maintained the test item in their working memory, he or she would always correctly identify the change item during test. Contrarily, if the participant failed to maintain the item in their visual working memory, they would randomly guess one out of the N total items they studied as the one that changed its relevant feature. Therefore, the accuracy under the two circumstances would be:

$$Acc(success) = P(success)*1 = \frac{K}{N}$$
$$Acc(fail) = P(fail)*\frac{1}{N} = \frac{(N-K)}{N2}$$

respectively.

Notice here that we adopted the random guessing rule that assumed participants would randomly select from the entire set of items in the test display when they had not successfully detected a change. Although the optimal strategy would be to avoid choosing any item that had been stored (Rhodes et al., 2018), the data were inconsistent with the application of this optimal strategy. Instead, pilot work showed that the
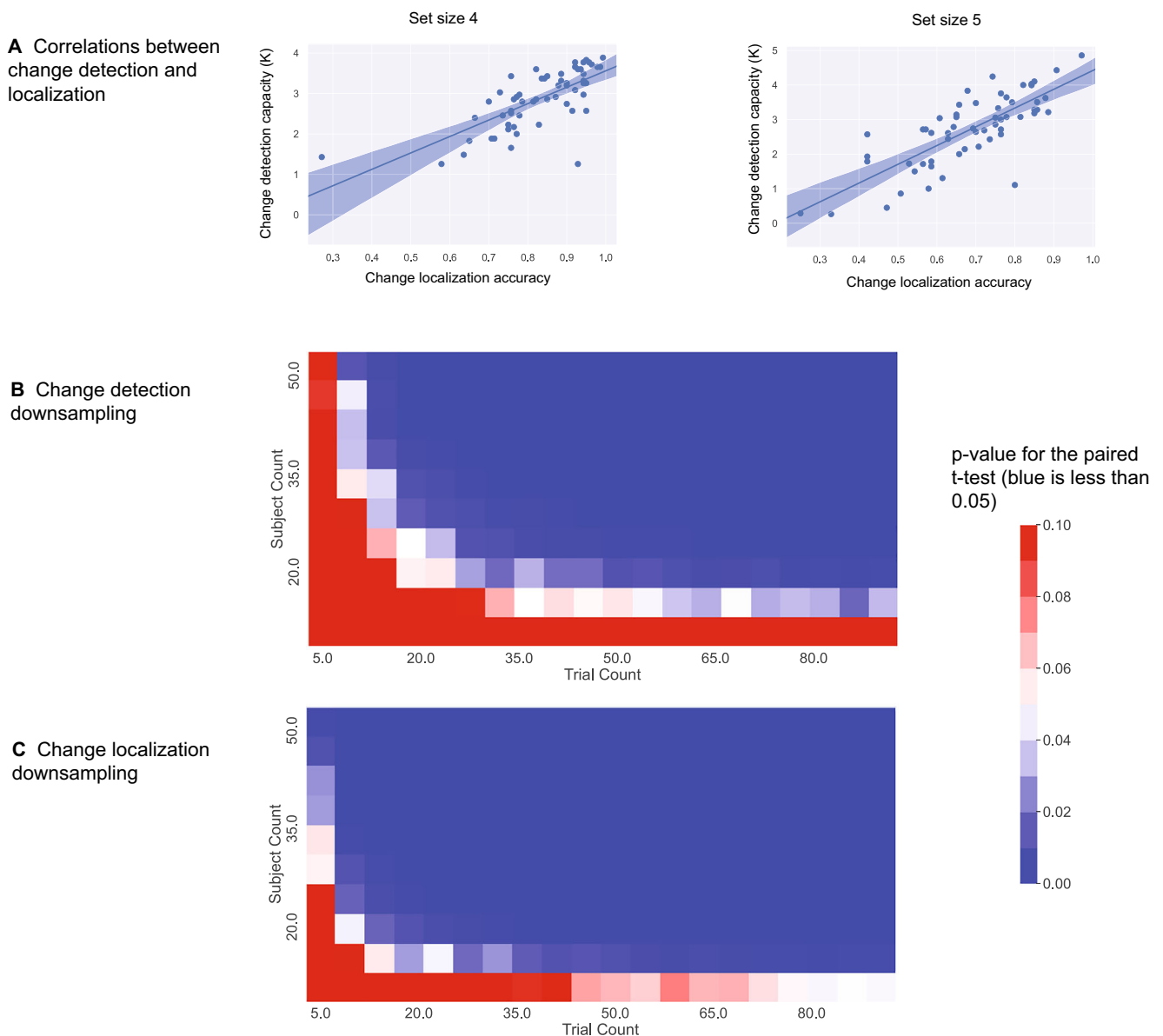
Fig. 7 Change detection and change localization paradigms captured similar individual differences under two smaller set sizes. **a** Change localization exhibited similar individual differences measured by change detection paradigm under set size 4 and 5. **b** Iterative downsampling of the p value resulted from the paired t-test on accuracies between set size 4 and 5 using change detection data. **c** Iterative downsampling of the p value resulted from the paired t-test on accuracies between set size 4 and 5 using change localization data

assumption of random guessing produced K scores that better matched estimates from change detection tasks in the same participants. Numerically, random guessing models constantly overestimate K scores compared to change detection (set size 4: 0.2, set size 5: 0.3), while informed guessing models constantly underestimate K scores (set size 4: -0.5, set size 5: -0.3). Therefore, accuracy in the change localization paradigm can be expressed as follows:

$$Acc(change\ localization) = Acc(success) + Acc(fail)$$

$$= \frac{(N-K)}{N^2} + \frac{K}{N}$$

Given that the set size N was a known variable for each participant, we could therefore derive K, the visual working memory capacity, from the accuracy of change localization:

$$K = \frac{Acc(change\ localization)*N^2 - N}{N-1}$$

In Experiment 3, each participant had their two K capacity estimates with set size 4 and 5 from the change detection paradigm, as well as two change localization accuracies with the same set sizes. The within-subject change localization and change detection design allowed us to calculate K capacity

from change localization task and validate with K capacity from the canonical change detection task.

First, the calculated change localization K was highly correlated to the change detection K scores for both set size 4 ($r(56) = 0.71$, $p = 2.91$x 10^{-10}) and set size 5 ($r(56) = 0.78$, $p = 4.97$ x 10^{-13}). As shown in Fig. 8, the mean difference between change localization K and change detection was small. The equation overestimated 0.22 squares on average out of an array with four items being studied and overestimated 0.32 squares out of a study array of five items, both reasonably small from the change detection K capacity estimates. Furthermore, both distributions were roughly symmetric around zero. We concluded that this equation could be effectively used in converting accuracy in change localization into K capacity score.

Discussion

In Experiment 3, we replicated previous work in showing that the change localization paradigm was statistically reliable with slightly smaller set sizes. Furthermore, our results revealed that the change localization paradigm was more sensitive in detecting the decline in working memory accuracy with increasing set sizes than the change detection paradigm. Lastly, we derived an equation that converted change localization accuracy into K, the number of items retained in visual working memory. Adopting a random guessing strategy, we showed that K measured by change localization was highly correlated to, and numerically similar to, K measured by the change detection paradigm.

In Experiment 4, we examined whether change localization would be a more sensitive procedure for detecting another established empirical pattern in the working memory literature. Estimates of working memory capacity (K) are intended to represent the total number of items a person can retain on average in these tasks (Cusack et al., 2009). Interestingly, this estimate has been shown to decline when participants are presented with larger sample arrays, such that estimates of K

decline as set size increases beyond about three items. This "overload" cost has been shown to be amplified in participants with lower working memory capacity (Balaban et al., 2019), such that there is a negative correlation between the size of the overload effect and separately acquired estimates of K. Here, we sought to document this empirical pattern – the correlation between WM capacity and the size of the overload effect – using both change detection and change localization. Since the overall reliability is expected to be lower for this difference score, the "overload" cost, compared to the measure of K, this experiment provided a further opportunity to examine whether change localization would provide a more sensitive procedure for documenting the relationship between working memory capacity and the overload effect.

## Experiment 4

Method

Participants

Two separate groups, each of 50 participants, were recruited at the rate of $9.50 per hour from Prolific, an online platform for participant recruitment. All participants were 18–35 years old, currently lived in the USA, had normal or corrected-to-normal vision, and had no ongoing psychological or neurological disorders.

Stimuli

The stimuli used were the same as in Experiment 1.

Procedures

For the first group of participants, we administered a color version of change detection task similar to those in Experiment 1. Each participant was asked to determine if the color of the
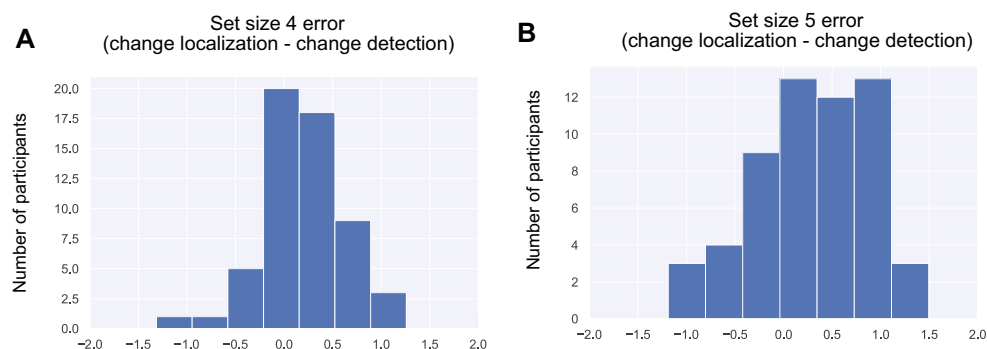
Fig. 8 K capacity derived from change localization highly resembled K measured in change detection. **a** Applying the equation of change localization K. the error between change localization K and change detection K were close to zero on average and symmetric around zero with set size 4. **b** The error between two Ks were also close to zero when the equation was applied on set size 5 accuracy data

probed item changed with four, six, or eight items during encoding. Each participant finished 120 trials of change detection for each set size, resulting in 360 trials in total.

For the second group of participants, each participant finished the color change localization task with four, six, or eight test items on screen. Each participant finished 120 trials of change detection for each set size, resulting in 360 trials in total.

Result

Change localization was more sensitive to the overload effect than the change detection paradigm

In the change detection task, we observed the typical pattern that the participants on average had stored less items when presented with set size 8 arrays compared to set size 4 arrays ($t(49) = 2.54$, $p = 0.01$, see Fig. 9a). The change localization paradigm also showed a higher measured capacity under set size 4 than set size 8 ($t(49) = 4.38$, $p = 6.15 \times 10^{-5}$, see Fig. 9c).

Another iterative down-sampling analysis revealed that change localization could reliably detect the overload effect with fewer trials and subjects. We performed iterative down-sampling on the t score comparing K capacity under set size 4 and set size 8 for the two paradigms. In our experiment, we adopted 100 iterations of runs in the down-sampling procedure, meaning that each number of trials and subjects were

randomly sampled for 100 times from our dataset. We randomly sampled the number of subjects from five to 50 in steps of five, and the number of trials from five to 120 in steps of five. As Fig. 9 illustrates, change localization required a smaller number of trials and subjects to observe a significant drop in K from set size 4 to 8, due to the pattern of the green squares being more left and lower compared to the change detection figure (see Fig. 9b and d).

Change localization was more sensitive to the correlation between K and the size of the overload effect

In the change detection task, we replicated past findings that higher K subjects showed a smaller drop in K when set size was increased. To document this effect in the present study, we measured the correlation between set size 6 K estimates and the difference between K estimates under set size 4 and 8. We saw a negative correlation between K and the drop between set size 4 and 8, meaning that high-capacity subjects exhibited a smaller overload effect than lower capacity subjects ($r(49) = 0.42$, $p = 0.0025$, see Fig. 10a). The change localization paradigm also showed the same negative correlation ($r(49) = -0.54$, $p = 4.68 \times 10^{-5}$, see Fig. 10c).

Given that both paradigms showed that K capacity was negatively correlated to the capacity drop with larger set sizes, we further examined which paradigm could reliably detect this
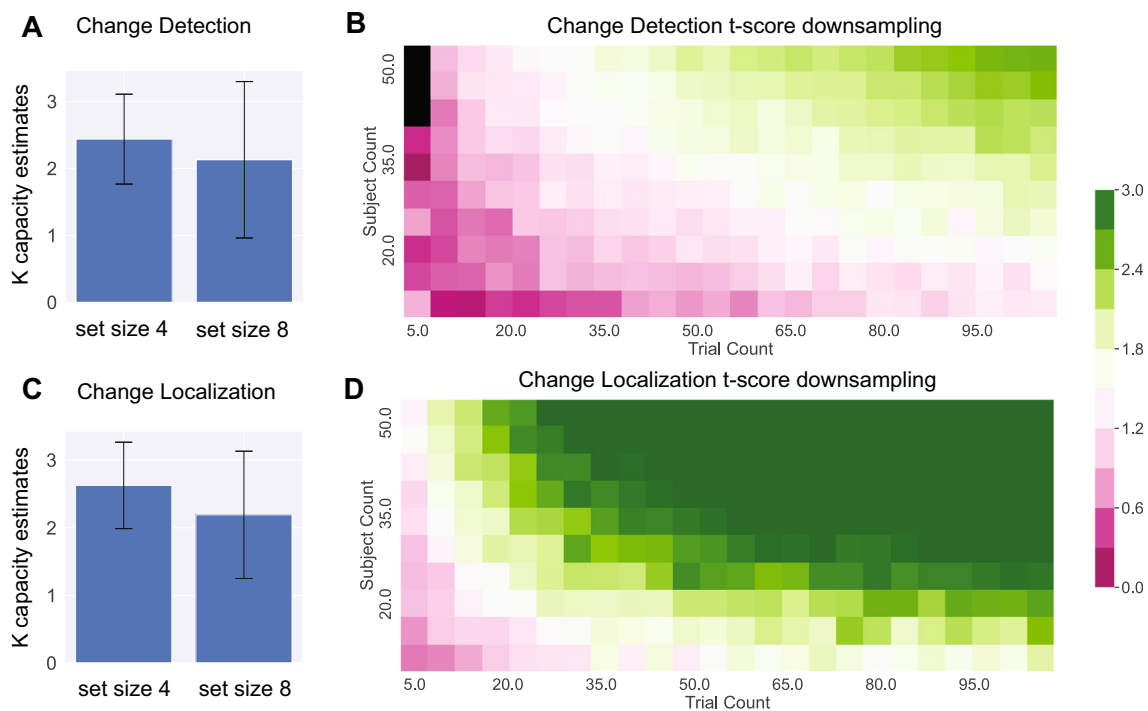
Fig. 9 Change localization was statistically more sensitive in capturing set size effect than change detection paradigm. **a** Change detection paradigm showed signfiicant drop in K estimate from set size 4 to set size 8. **b** Iterative downsampling of z-score comparing K estimates under set size 4 condition to that of set size 8 condition in change detection data.

c Change localization paradigm also showed signfiicant drop in K estimate from set size 4 to set size 8. **d** Iterative downsampling of z-score comparing K estimates under set size 4 condition to that of set size 8 condition in change localization data
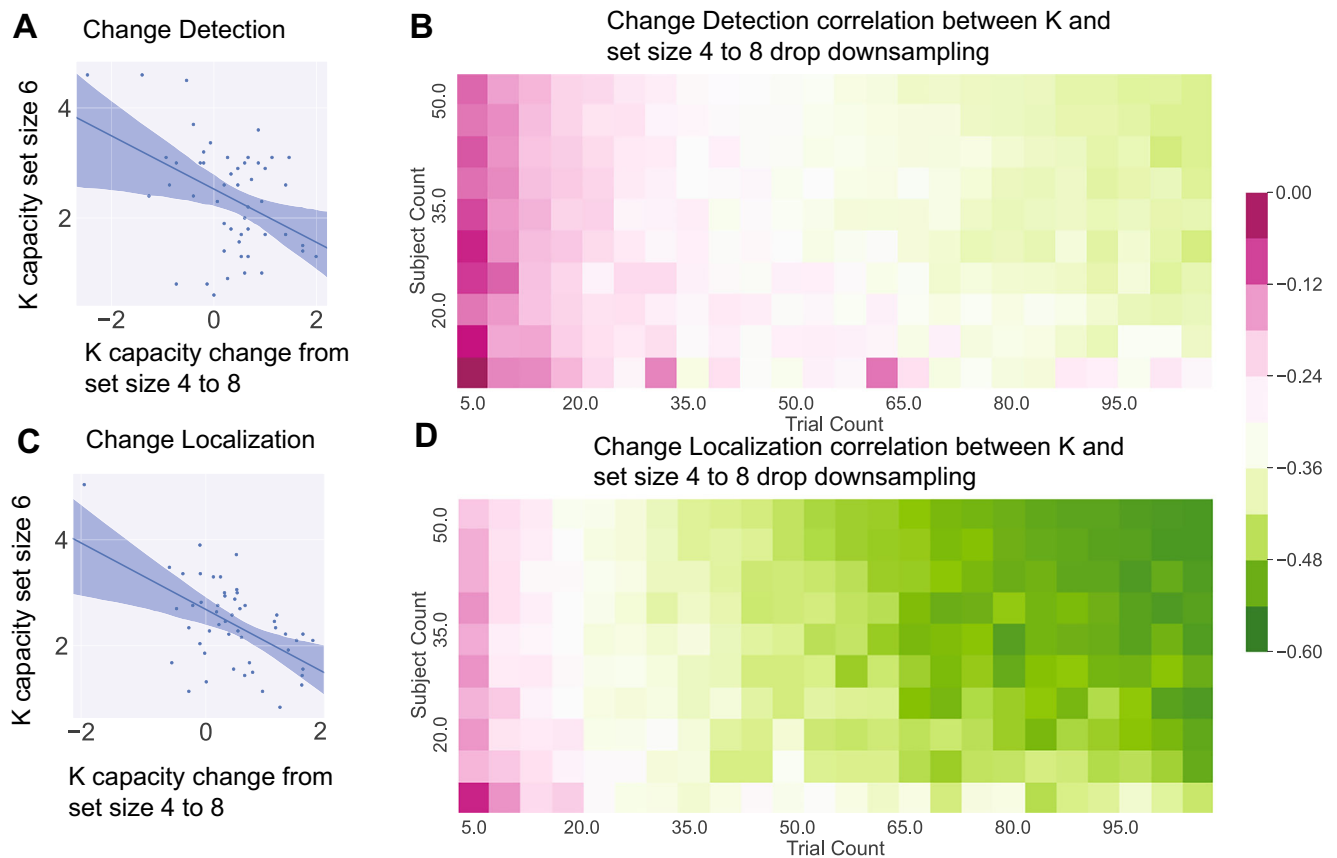
Fig. 10 Change localization was statistically more sensitive in capturing the set size effect that high capacity subjects would improve more when being tested with 8 stimuli compared to 4 stimuli than the change detection paradigm. **a** Change detection paradigm showed that K capacity estimated with a left-out set size 6 task negatively correlated to the K capacity change from set size 4 to set size 8. **b** Iterative downsampling of r correlations between K estimates and the change in K from set size 4 condition to set size 8 condition in change detection data. **c** Change localization paradigm also showed that K capacity estimated with a left-out set size 6 task negatively correlated to the K capacity change from set size 4 to set size 8. **d** Iterative downsampling of r correlations between K estimates and the change in K from set size 4 condition to set size 8 condition in change localization data

correlation with fewer trials and subjects. We performed iterative down-sampling on the Pearson correlation between K capacity and the difference in capacity estimated by set size 4 and set size 8 for the two paradigms. In our experiment, we adopted 100 iterations of runs in the down-sampling procedure, meaning that each number of trials and subjects were randomly sampled for 100 times from our dataset. We randomly sampled the number of subjects from five to 50 in steps of five, and the number of trials from five to 120 in steps of five. Change localization required a smaller number of trials and subjects to observe a significant correlation between K and the drop in performance from set size 4 to 8, as illustrated by the broader swath of green squares in Fig. 10d (change localization) than in Fig. 10b (change detection).

Discussion

In Experiment 4, we replicated previous findings of an "overload" effect in which capacity estimates decline as set sizes are increased rather beyond capacity limits. Furthermore, we

replicated the finding that this overload effect was negatively correlated with working memory capacity, such that higher K subjects showed a smaller decline in performance with larger set sizes. The change localization paradigm showed both effects, similar in magnitude to the change detection paradigm. Lastly, the change localization paradigm required fewer trials and subjects to observe similar empirical patterns for both effects compared to change detection paradigm. These findings provide further evidence that change localization may be superior to change detection in terms of both reliability and sensitivity to known empirical patterns.

General discussion

Here we show that change localization provides a highly reliable and sensitive method for measuring variations in working memory capacity across individuals and experimental conditions. Participants viewed a whole-field test display and selected the item that had changed compared to the sample

display. We showed that this new paradigm tapped into the same variance across individuals as the change detection task, using three different set sizes and both color and shape stimuli. Furthermore, the new paradigm required less than half of the number of trials (about 50–75 trials) to achieve strong reliability as a measure of individual differences. Lastly, we confirmed that change localization was sensitive to two well-known effects in the change detection literature, and the new paradigm required far fewer trials to reveal these effects.

Change localization may be superior to change detection in part because of a substantially lower level of chance performance for localization (1 in 6) versus detection (1 in 2). Lower chance level in change localization enabled higher variance across individuals who had above-chance performance. Additionally, the smaller number of trials change localization needed made it a faster test than change detection. Thus, in addition to providing a measure with stronger psychometric properties, it will also be an expedient approach to use whenever data collection time is a limiting factor, such as with special populations. Additionally, a change localization paradigm may provide a useful alternative in a number of other domains, such studies employing scene (Caplovitz et al., 2008) or video (Levin et al., 2018) stimuli.

Another virtue of the change localization measure is that it is bias-free, because participants have no opportunity to set a response threshold (whether liberal or conservative) for whether a change was detected or not. Indeed, Williams et al. (2022) have reported that when response bias is very strong, change detection scores may underestimate true memory capacity. Thus, bias-free measures like ours (e.g., Adam et al., 2015; Cowan et al., 2012; Eng et al., 2005) avoid this potential source of error. That said, we also examined whether our data provided evidence for the impact of response bias in change detection that was described by Williams et al. First, we found that capacity estimates using change localization were reliably higher than estimates obtained with change detection. To examine whether this increase was linked with the degree of response bias in change detection, we examined whether the size of this advantage for change localization was related to a behavioral measure of response bias (i.e., a larger proportion of "no-change" responses was taken as evidence of a more conservative response bias). Across three experiments, we did not find evidence that response bias was explained by the conservativeness of participants (r = 0.13, -0.03, 0.25, respectively; p = 0.20, 0.82, 0.06, respectively). Therefore, while Williams et al. (2022) are correct that a strong response bias can yield an artificial decline in capacity estimates, our large sample did not show the predicted relationship between response bias and WM capacity estimates. Moreover, while it is likely that response bias affects performance in change detection tasks, the high convergent validity between change detection estimates of WM capacity and a wide array of cognitive abilities (e.g., Fukuda et al., 2010;

Unsworth et al., 2014) suggests that a substantial proportion of between-subject variance in change detection is not attributable to response bias. Finally, change detection has been shown to be a robust method for documenting the effects various experimental manipulations such as the number of items stored (Luck & Vogel, 1997), the presence of distractors (Vogel et al., 2005), and the impact of statistical regularities (Umemoto et al., 2010). The presence of a response bias does not undermine these uses of this metric for WM capacity.

To conclude, we have shown that change localization is a highly reliable and sensitive approach for measuring WM capacity. Change localization requires only half as many trials as change detection to achieve robust reliability, and provides a more sensitive measure for efficiently documenting known empirical patterns from the change detection literature. Our hope is that this may be a useful approach for a wide range of questions within regarding this core aspect of cognitive ability.

Open science statement The data, code, and materials for all experiments will be publicly accessible on OSF upon publication. There is not a preregistration for our studies here.

References

Adam, K. C., Mance, I., Fukuda, K., & Vogel, E. K. (2015). The contribution of attentional lapses to individual differences in visual working memory capacity. *Journal of Cognitive Neuroscience, 27*(8), 1601–1616.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science, 18*(7), 622–628.

Balaban, H., Fukuda, K., & Luria, R. (2019). What can half a million change detection trials tell us about visual working memory? *Cognition, 191*(May), 103984. https://doi.org/10.1016/j.cognition.2019.05.021

Caplovitz, G. P., Fendrich, R., & Hughes, H. C. (2008). Failures to see: Attentive blank stares revealed by change blindness. *Consciousness and Cognition, 17*(3), 877–886. https://doi.org/10.1016/j.concog.2007.08.006

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114.

Cowan, N., Rouder, J. N., Blume, C. L., & Scott Saults, J. (2012). Models of verbal working memory capacity: What does it take to make them work? *Psychological Review, 119*(3), 480–499. https://doi.org/10.1037/a0027791

Cusack, R., Lehmann, M., Veldsman, M., & Mitchell, D. J. (2009). Encoding strategy and not visual working memory capacity correlates with intelligence. *Psychonomic Bulletin and Review, 16*(4), 641–647. https://doi.org/10.3758/PBR.16.4.641

Eng, H. Y., Chen, D., & Jiang, Y. (2005). Visual working memory for simple and complex visual stimuli. *Psychonomic Bulletin & Review, 12*(6), 1127–1133.

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*(5), 673–679.

Levin, D. T., Seiffert, A. E., Cho, S. J., & Carter, K. E. (2018). Are failures to look, to represent, or to learn associated with change blindness during screen-capture video learning? *Cognitive Research: Principles and Implications, 3*(1). https://doi.org/10.1186/s41235-018-0142-3

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279–284. https://doi.org/10.1038/36846

Miller, A. L., Gross, M. P., & Unsworth, N. (2019). Individual differences in working memory capacity and long-term memory: The influence of intensity of attention to items at encoding as measured by pupil dilation. *Journal of Memory and Language, 104*, 25–42. https://doi.org/10.1016/j.jml.2018.09.005

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 and Cognition, 43*(3), 397–420. https://doi.org/10.3758/s13421-014-0492-0

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics, 44*(4), 369–378.

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences, 20*(2), 110–122. https://doi.org/10.1016/j.lindif.2009.10.005

Rhodes, S., Cowan, N., Hardman, K. O., & Logie, R. H. (2018). Informed Guessing in change detection. *Journal of Experimental Psychology: Learning Memory and Cognition, 44*(7), 1023–1035. https://doi.org/10.1037/xlm0000495

Umemoto, A., Scolari, M., Vogel, E. K., & Awh, E. (2010). Statistical learning induces discrete shifts in the allocation of working memory resources. *Journal of Experimental Psychology: Human Perception and Performance, 36*(6), 1419.

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. https://doi.org/10.1016/j.cogpsych.2014.01.003

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature, 438*(7067), 500–503.

Williams, J., Robinson, M. M., Schurgin, M., Wixted, J., & Brady, T. (2022). You can't "count" how many items people remember in visual working memory: The importance of signal detection-based measures for understanding change detection performance. https://doi.org/10.31234/osf.io/jnt7r

Xu, Z., Adam, K. C. S., Fang, X., & Vogel, E. K. (2018). The reliability and stability of visual working memory capacity. *Behavior Research Methods, 50*(2), 576–588. https://doi.org/10.3758/s13428-017-0886-6