Cognition 191 (2019) 103984

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

What can half a million change detection trials tell us about visual working memory?

Halely Balaban[a,*], Keisuke Fukuda[b], Roy Luria[a]

[a] Sagol School of Neuroscience and The School of Psychological Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[b] Department of Psychology, University of Toronto Mississauga, Mississauga, ON L5L 1C6, Canada

ARTICLE INFO

Keywords:
Working memory capacity
Individual differences
Change detection

ABSTRACT

Visual working memory (VWM) represents the surrounding world in an active and accessible state, but its capacity is severely limited. To better understand VWM and its limits, we collected data from over 3,800 participants in the canonical change detection task. This unique population-level data-set sheds new light on classic debates regarding VWM capacity. First, the result supported a view of VWM as an active process, as manifested by the fact that capacity estimates were not stable across set-sizes, but rather lower for the larger set-size. Another support for this notion came from the tight connection capacity estimates had with a measure of attentional control. Together, the data suggested that individual differences in capacity do not reflect only differences in storage-size, but differences in the efficiency of using this storage. Second, we found a response bias such that subjects are more likely to respond that the probed item changed, and this criterion bias was further shifted as the set-size increased. These findings are naturally explained by a slot-like theory arguing that when load exceeds capacity, certain items remain completely outside of VWM (instead of all items being represented in lower resolution), therefore causing subjects to perceive them as different from VWM contents even when they are unchanged. Additionally, we found that the pattern of d' also confirmed the predictions of a slot-like view of VWM, such that some items are represented with high fixed resolution and others are not represented at all, although this finding is based on two measures with very different underlying assumptions. We also discuss how flexible-resource views can accommodate these results. Moreover, comparing performance between the first and last trials demonstrated no evidence for proactive interference as the driving factor of capacity limitations. We provide further details regarding the distribution of individual capacity, the relations between capacity and demographic variables, and the spatial prioritization of the items.

1. Introduction

Visual working memory (VWM) holds relevant visual information in an active state, ready to be accessed and manipulated by higher cognitive functions (Cowan, 2001). However, only a very limited amount of information can be retained in this privileged state, creating a bottleneck for how we process incoming information. Corroborating the importance of VWM in everyday behavior, it is specifically damaged in a range of conditions such as Alzheimer's disease, attention deficit hyperactivity disorder (ADHD), old age, and schizophrenia (e.g., Johnson et al., 2013; Jost, Bryck, Vogel, & Mayr, 2011; Martinussen, Hayden, Hogg-Johnson, & Tannock, 2005; Parra et al., 2011). The nature of VWM capacity is heavily debated (e.g., whether it is better described as a continuous or discrete resource, cf. Brady, Konkle, & Alvarez, 2011; Luck & Vogel, 2013; Ma, Husain, & Bays, 2014), but the existence of a severe limitation on this capacity is vastly acceptable and can be considered as one of the defining characteristics of VWM.

Despite average capacity limits being quite low (~3 simple items' worth of information), a great deal of variability exists between individuals. Capacity is highly stable at the individual level (e.g., across set-sizes or blocks, Cronbach's alphas > 0.9; Xu, Adam, Fang, & Vogel, 2018), and is tightly correlated with measures of fluid intelligence, attentional control, and many aptitude measures (e.g., Cowan et al., 2005; Fukuda, Vogel, Mayr, & Awh, 2010; Vogel, McCollough, & Machizawa, 2005). Understanding the nature of these capacity differences and further exploring how VWM is related to other important cognitive constructs is the focus of numerous ongoing research projects.

A prominent method for quantifying VWM performance is the change detection paradigm (Luck & Vogel, 1997; Pashler, 1988; Phillips, 1974). In the canonical form of the task, several simple items

* Corresponding author.
E-mail address: halelyba@mail.tau.ac.il (H. Balaban).

https://doi.org/10.1016/j.cognition.2019.05.021
Received 9 November 2018; Received in revised form 16 May 2019; Accepted 25 May 2019
Available online 21 June 2019

are briefly presented, and after a short retention interval the test array appears, with an item in one of the previously-occupied locations. The subjects' task is to indicate whether the probed item, i.e., the item in the test array, is the same or different (typically with an equal probability) from the item appearing in the same place in the memory array. Performance is usually transformed into a capacity estimate (K), i.e., the number of items that can be successfully retained in VWM (Cowan, 2001; Pashler, 1988). Even a ~10 min long version of this task is highly reliable, with capacity measures being stable between testing sessions more than a year apart (r = 0.77; Johnson et al., 2013). This makes the paradigm ideal for examining individual differences in capacity.

In the present study, we report the results of over 3,800 subjects that completed a short change detection task in our lab, before participating in one of many different experiments. Usually, each experiment (even when sample-size is adequate to examine individual differences) provides a small number of trials overall. Pooling together trials from an extremely large number of subjects, for a total of over 460,000 trials of the same task, allowed us to analyze the data in ways that are beyond the scope of ordinary studies. Our main goal was to shed new light on VWM capacity limitations. Additionally, the large sample size allowed us to better characterize the change detection task, in terms of performance biases, both at the individual and at the "population" level.

An important advantage of the canonical change detection paradigm is that task performance is only minimally influenced by processes external to VWM, such as verbal working memory, iconic memory, long-term memory, and response-related processes (Cowan, 2001). This had made change detection performance almost synonym to VWM capacity. However, it is important to remember that the change detection task is not a cognitive process per se, but a paradigm, and the ability of this paradigm to adequately measure VWM processes depends on many carefully-thought parameters of the experimental setup (Vogel, Woodman, & Luck, 2001). Specifically, isolating VWM processes requires using short presentation times or adding articulatory suppression to prevent verbal coding (Luck & Vogel, 1997; Vogel et al., 2001), including masking or a long-enough (~1 s) retention interval to delete the retinal after-image (Phillips, 1974), and using simple and highly discriminable items to minimize comparison-process errors that arise for complex stimuli (Awh, Barton, & Vogel, 2007). Adhering to these conventions allow strong conclusions about VWM to be drawn directly from change detection performance.

2. Materials and methods

All analyzed data-sets can be found at the Open Science Framework: http://dx.doi.org/10.17605/OSF.IO/MZS9E.

2.1. Participants

Subjects were mostly Tel Aviv University students and several individuals from the Tel Aviv University community, who completed the task before taking part in a longer experiment (a small subset of the K estimates data was previously published in Allon & Luria, 2017; Allon, Vixman, & Luria, 2018; Vaskevich & Luria, 2018). Subjects participated for either payment (approximately 40 NIS, or $10, per hour), partial course credit, or voluntarily without compensation. All had normal or corrected-to-normal visual acuity and normal color vision. All participants gave informed consent following the procedures of protocols approved by the Ethics Committee at Tel Aviv University. Age ranged between 18 and 39.

A total of 3,923 data-sets were collected for the following analyses. Some participants took part in more than one experiment at the lab, and thus contributed more than one data-set to the current analysis (approximately 1,000 data sets came from such repeated participants). We excluded data-sets with below-chance performance at set-size 4 or a K estimate of less than 0 (46 data-sets, 1.2%), because this results from below-chance performance which likely reflects switching the response keys (i.e., pressing the "same" key for a "different" response and vice versa). Additionally, due to an error in the code, some participants completed more than 120 trials. We included data-sets with 121–130 trials (54 data-sets, 1.4%), and excluded data-sets with over 130 trials (28 data-sets, 0.7%). This left a total of 3,849 analyzed data-sets with 462,186 trials.

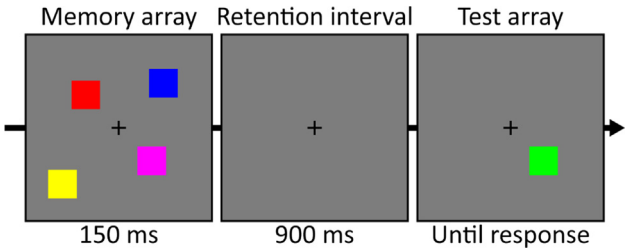

Fig. 1. Trial sequence in the change detection task employed in the present study. A memory array of colored squares appeared for a short duration, followed by a blank retention interval, and then a test array with only one probe that can be either the same as the item in the same location in the memory array, or different. This is an example of a set-size 4 trial, including a change.

2.2. Stimuli and procedure

Subjects performed a ~10 min color change detection task (see Fig. 1). A memory array of colored squares (approximately 1.3° × 1.3° of visual angle, from a viewing distance of approximately 60 cm) appeared for 150 ms around a black fixation cross (0.4° × 0.4°) at the center of a grey (RGB values: 125,125,125) screen. Items were randomly selected, without replacement, from a set of 9 highly discriminable colors: red, magenta, blue, cyan, green, yellow, orange, brown, and black (RGB values, respectively: 254,0,0; 255,0,254; 0,0,254; 0,255,255; 0,255,1; 255,255,0; 255,128,65; 128,64,0; 0,0,0; note that the colors would render differently depending on the video card and monitor). Participants were instructed to memorize the squares' colors. The squares disappeared for a 900 ms retention interval (leaving only the fixation cross), and then the probe appeared: a single square in one of the previously occupied locations (randomly determined). Participants' task was to indicate in a non-speeded manner, via button press, whether the probe was the same color as the square in that location, or a different color (using the "Z" and "/" keys on a standard keyboard, respectively; a subset of the subjects did a counterbalanced version, see the Results section).

Half of the trials included 4 colors in the memory array, and the other half included 8 colors (randomly intermixed). Half of the trials included a change, and half did not (randomly intermixed). After about 6 practice trials, participants completed one block of 120 experimental trials, with 30 trials for each combination of set-size (4 or 8) and trial type (change or no-change).

On each trial, one quarter of the items (1 in set-size 4 trials and 2 in set-size 8 trials) were presented on each quadrant. Items' locations were selected randomly from 9 possible locations in each quadrant: 2.7°, 5.3°, or 8° vertically from the center of the screen, and 2.7°, 5.3°, or 8° horizontally from the center of the screen (e.g., in the top left quadrant, items could appear centered at 2.7°, 5.3°, or 8° to the left of the fixation, and 2.7°, 5.3°, or 8° above the fixation).

2.3. Individual-level analyses

We extracted from each data-set a measure of K, using a standard formula (Cowan, 2001; Pashler, 1988): $K = N \times (H - FA)$, where K is the capacity estimate, N is the set-size, H is hit rate, and FA is the false alarm rate (meaning the individual proportion of correct responses in change trials, and of incorrect responses in no-change trials, respectively). In the present situation, of a single-probe task with a 50%

probability of change, an equivalent formula is $K = N \times (2 \times Accuracy - 1)$, where *Accuracy* is the proportion correct across all trials of that set-size (regardless of whether there was or wasn't a change). Following standard practice, we averaged K across the two set-sizes.

Additionally, we calculated the signal detection theory measures of d', which in the present context is an index of memory signal strength, and criterion, following standard practices for these kinds of "yes/no" tasks (Macmillan & Creelman, 2004). Note that in the change detection paradigm, since the task is to indicate whether a change occurred, a "yes" response corresponds to a changed or new item, unlike in classical "old/new" recognition task in which a "yes" response corresponds to a previously-shown item (e.g., Keshvari, van den Berg, & Ma, 2013). Memory signal strength, i.e., d', was calculated as: $d' = Z(H) - Z(FA)$, where $Z(H)$ is the Z-score of the hit rate, and $Z(FA)$ is the Z-score of the false alarm rate. Criterion, or bias, was calculated as $C = -0.5 \times (Z(H) + Z(FA))$, where C is the criterion, and $Z(H)$ and $Z(FA)$ are as in the d′ calculation.

2.4. Group-level analyses

To examine several fine-grained characteristics of the change detection task, we pooled together all data-sets, for a total of 462,186 trials. We then performed two analyses on the aggregate set of trials. First, we computed K separately for each serial position of the trial within the experiment (1–120; 3,849 trials per serial position number). This means we treated all trials from a single position as if they came from a single subject that performed 3,849 trials and then extracted K in the usual way. Second, we computed accuracy separately for each of the 36 probed locations (12,625–13,039 trials per location). For this analysis, we also report a subset of 905 participants who did a counterbalanced version of the task in terms of the response keys mapping, in which we break down trials also by the type of trial (change vs. no-change).

2.5. Statistical analyses

Due to the very large sample-size in most of our analyses, traditional statistical tests are not informative, because they're likely to be significant even for trivial effects. For the same reason, traditional confidence intervals (CIs) will be extremely narrow and hence not indicative. Instead, for our main analyses we relied on the large sample-size and treated the data-sets as a population. We report the observed 2.5 and 97.5 percentiles, i.e., the values that mark the range holding the central 95% of the "population" distribution. We refer to these observed percentiles as the *population 95% range*. We additionally report 95% Bayesian credibility intervals, calculated using the JASP software (with default priors). We report traditional 95% CIs for correlations measures, because for these measures a single value is computed for all subjects. To test whether the data is normally distributed, we performed D'Agostino's K^2 test (D'agostino, Belanger, & D'agostino, 1990), based on skewness and kurtosis, which is suitable for our large sample-size.

3. Results

3.1. Individual VWM capacity estimates

For each individual data-set, we calculated a measure of K, according to the above-mentioned formula. Average K was 2.66, with a standard deviation of 0.83. Table 1 summarizes descriptive statistics for all "population-level" reported measures.

This replicates two of the key characteristics regarding VWM capacity. First, the highly limited nature of this workspace (less than 3 simple items) is very similar to previous reports, including the findings of a recent large-scale study conducted in the US (average K: 2.55, N = 495; Fukuda, Woodman, & Vogel, 2015), though somewhat higher than the findings of one conducted in China (average K: 2.14, N = 135; Xu et al., 2018). Second, we found large individual differences, as manifested by the 0.8 items standard deviation. Fig. 2 depicts the distribution of K values across participants.

As can be seen in Fig. 2, the observed K values appear to follow a normal distribution. To qualitatively estimate this, we present a Q-Q plot in Fig. 3. This plot compares our observed quantiles to the theoretical quantiles expected from a normal distribution. The more similar the two compared distributions, the closer the individual data points will be to a straight line. As can be seen from the figure, the observed data indeed closely matches the expected quantiles, suggesting that the data is normally distributed. To quantitively support this observation, we performed the D'Agostino K^2 test, which suggested that the distribution is not significantly different from a normal distribution ($K^2 = 1.23$, $p = 0.54$). It is noteworthy that the results of an extremely simple 10-min color memory task are normally distributed.

It might be argued that this normality does not reflect the shape of the K construct itself, but rather the normality of the measurement error in this task, resulting from a sum of binary choices. However, K estimates for set-size 8 separately was not normally distributed ($K^2 = 31.08$, $p < 0.001$; we tested this for set-size 8 because although K at set-size 4 was also not normally distributed, it has a relatively low upper bound), despite also involving similar measurement errors. In fact, K estimates for 8 items were positively skewed, and K estimates for 4 items were negatively skewed, which suggests a different alternative explanation for the normality of average K estimates: the normal distribution might be an averaging artifact. To rule this out, we averaged for each individual data-set two other oppositely-skewed measures, namely d' which was positively skewed and criterion which was negatively skewed (we used the measures from trials of both set-sizes combined). We found that the average measure was not normally distributed ($K^2 = 26.91$, $p < 0.001$), meaning that not every two oppositely-skewed measures will create a normal distribution when averaged together. Finally, it might be argued that it is the variance in individual estimates of K that follows a normal distribution (i.e., measuring each participant numerous times would have produced a normal distribution of K estimates for that individual), thus producing a normal distribution across participants. If this is the case, however, it should also hold for other measures. We tested this for d' and criterion, and found that they were not normally distributed, either when considering each set-sizes separately or when aggregating both set-sizes (all K^2s > 18, all p's < 0.001). Thus, the normal distribution of K estimates is an important finding, which we suggest reflects the underlying normal distribution of VWM capacity in the population, demonstrating that K is a meaningful construct.

Next, we examined K estimates at the different set-sizes. Capacity estimates at set-size 4 trials were higher than in set-size 8 trials (mean K: 2.79 vs. 2.54, respectively), and varied less (SD: 0.66 vs. 1.21, respectively). Since both set-sizes are beyond the average capacity limit (i.e., most participants cannot hold even 4 simple items in VWM), the fact that the larger set-size produces a lower capacity estimates is interesting. A view of VWM as a passive storage would lead to a prediction that capacity should be stable across set-size (e.g., if one can only hold 2 items, they should always hold 2 items, regardless of the presented set-size). This is not the case, however, and increasing the set-size further decreases performance, which strongly indicates that VWM is a dynamic process. Thus, capacity estimates do not solely reflect the size of the storage, but also the ability to use it well under different circumstances, for example varying memory loads.

If that is the case, we should expect that average capacity estimates mainly reflect the larger set-size, which poses a greater difficulty. To test this, we examined the correlations between K estimates at each set-size separately, and overall K. K estimates at set-size 4 and 8 had a correlation of $r = 0.53$ (95% CI: [0.51, 0.56]), meaning they shared 28% of their variance. The correlation between overall K and set-size 4 K was $r = 0.79$ (95% CI: [0.77, 0.80]), while the correlation between overall K and set-size 8 K was $r = 0.94$ (95% CI: [0.94, 0.95]), see

Table 1
Descriptive statistics for the different measures: mean (standard deviation in parentheses), population 95% range (2.5 and 97.5 percentiles, see Methods), and lower and higher boundaries of the 95% Bayesian credibility interval (CI).

Measure	Mean (SD)	Population 95% range		95% Bayesian CI	
		2.5 percentile	97.5 percentile	Lower boundary	Higher boundary
K (averaged across set-sizes)	2.66 (0.83)	1.07	4.27	2.64	2.69
K: set-size 4	2.79 (0.66)	1.33	3.87	2.77	2.81
K: set-size 8	2.54 (1.21)	0.27	5.07	2.50	2.58
4–8 drop	0.25 (1.03)	−1.87	2.13	0.21	0.28
Hit rate	0.90 (0.08)	0.70	1.00	0.89	0.90
Correct rejection	0.61 (0.13)	0.33	0.85	0.61	0.62
Hit rate: set-size 4	0.93 (0.07)	0.77	1.00	0.93	0.93
Correct rejection: set-size 4	0.77 (0.14)	0.43	1.00	0.76	0.77
Hit rate: set-size 8	0.86 (0.11)	0.60	1.00	0.86	0.87
Correct rejection: set-size 8	0.46 (0.16)	0.17	0.77	0.45	0.46
Criterion: set-size 4	−0.41 (0.36)	−1.11	−0.28	−0.42	−0.40
Criterion: set-size 8	−0.69 (0.44)	−1.56	0.10	−0.71	−0.68
d': set-size 4	2.48	1.03	4.23	2.45	2.50
d': set-size 8	1.14	0.11	2.34	1.13	1.16
K: females	2.75 (0.81)	1.23	4.33	2.71	2.79
K: males	2.64 (0.82)	1.07	4.27	2.58	2.70

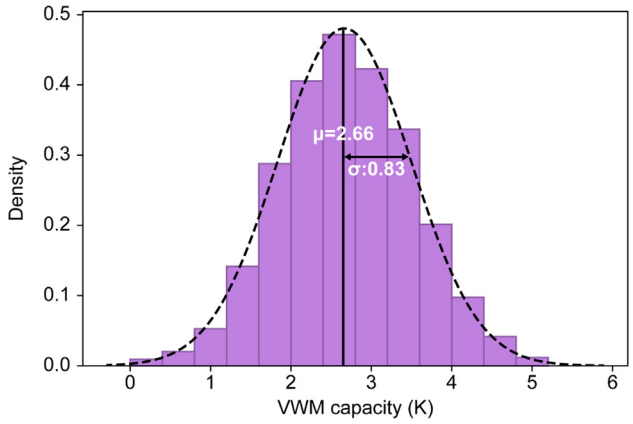

Fig. 2. A histogram of K (capacity estimate) values for our 3,849 individual data-sets (density on the Y axis). The dashed black line depicts a normal distribution with the same mean (μ) and standard deviation (σ) as our observed distribution, and these values are also presented on the histogram.

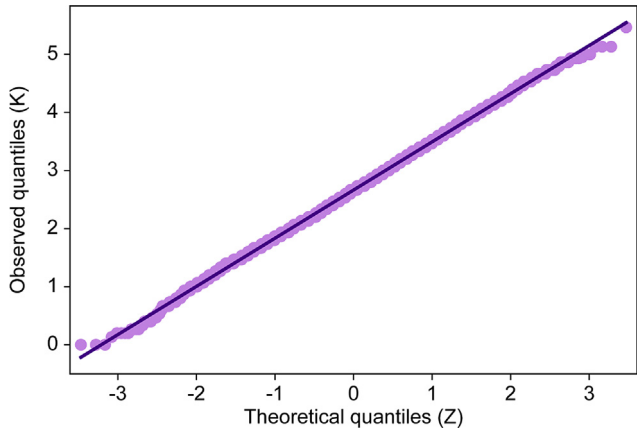

Fig. 3. A Q-Q plot of K values. Observed quantiles appear on the Y axis, and theoretical quantiles, derived from a normal distribution, appear on the X axis. The diagonal line depicts the standardized x = y line (i.e., the line on which the percentiles exactly match). The fewer deviations from the line, the more normally distributed the data is.

Fig. 4. Thus, the overall capacity estimate shared 62% of its variance with the set-size 4 estimate, and as much as 89% of its variance with the set-size 8 estimate. This tight relationship suggests that K mainly reflects the highest set-size (notably, because the variance is larger at set-size 8, its correlations are expected to be higher, but here it completely dominates the explained variance). This might lead one to argue that it could be enough to measure only set-size 8, because of its high correlation with overall K. Nevertheless, we argue that including only arrays that are far beyond average capacity is problematic, because their extreme difficulty might seriously damage subjects' motivation. We do note, however, that it is vital to include these set-sizes to properly estimate capacity.

The idea of VWM capacity as reflecting an active storage ability, instead of storage size, could explain why VWM capacity is correlated with the ability to filter out irrelevant distractors (Vogel et al., 2005): in tasks involving only relevant items, when the set-size exceeds the individual capacity limit, over-capacity targets are just like distractors. This means that set-size 8 is more difficult than set-size 4 partially because of the need to select a subset of items among more possibilities. For this reason, the difference between K at set-sizes 4 and 8 (termed "4–8 drop") was previously used as a measure of attentional control, argued to be conceptually separated from the size of VWM-storage (Unsworth, Fukuda, Awh, & Vogel, 2014, 2015). We found that the 4–8 drop was on average greater than zero (mean: 0.25), with substantial variance (SD: 1.03). While the 4–8 drop did not correlate with set-size 4 K ($r = 0.01$, 95% CI: [−0.02, 0.04]), it strongly correlated with set-size 8 K ($r = -0.84$, 95% CI: [−0.85, −0.83]) and with overall K ($r = -0.61$, 95% CI: [−0.63, −0.59]), such that high-capacity individuals had a smaller 4–8 drop, i.e., less of a difference between the two set-sizes (see Fig. 5). Because the 4–8 drop shares 37% of its variance with overall K and 71% of its variance with set-size 8 K, we argue that the 4–8 drop does not reflect an ability independent from capacity. Instead, both the 4–8 drop and classical capacity estimates seem to reflect, at least to some degree, the ability to flexibly use limited VWM resources. Another interesting possibility along this line is that the 4–8 drop reflects some sort of a strategy shift, which high-capacity individuals should be more likely to successfully utilize (as suggested by the correlations between VWM capacity and fluid intelligence, e.g., Fukuda et al., 2010). This also goes with a view of capacity as the efficient use of VWM.

Notably, the correlations were calculated between different measures derived from the same task, using the same small set of trials. This means that random factors that drive performance (e.g., attentional fluctuations or guesses) might affect several different measures in a

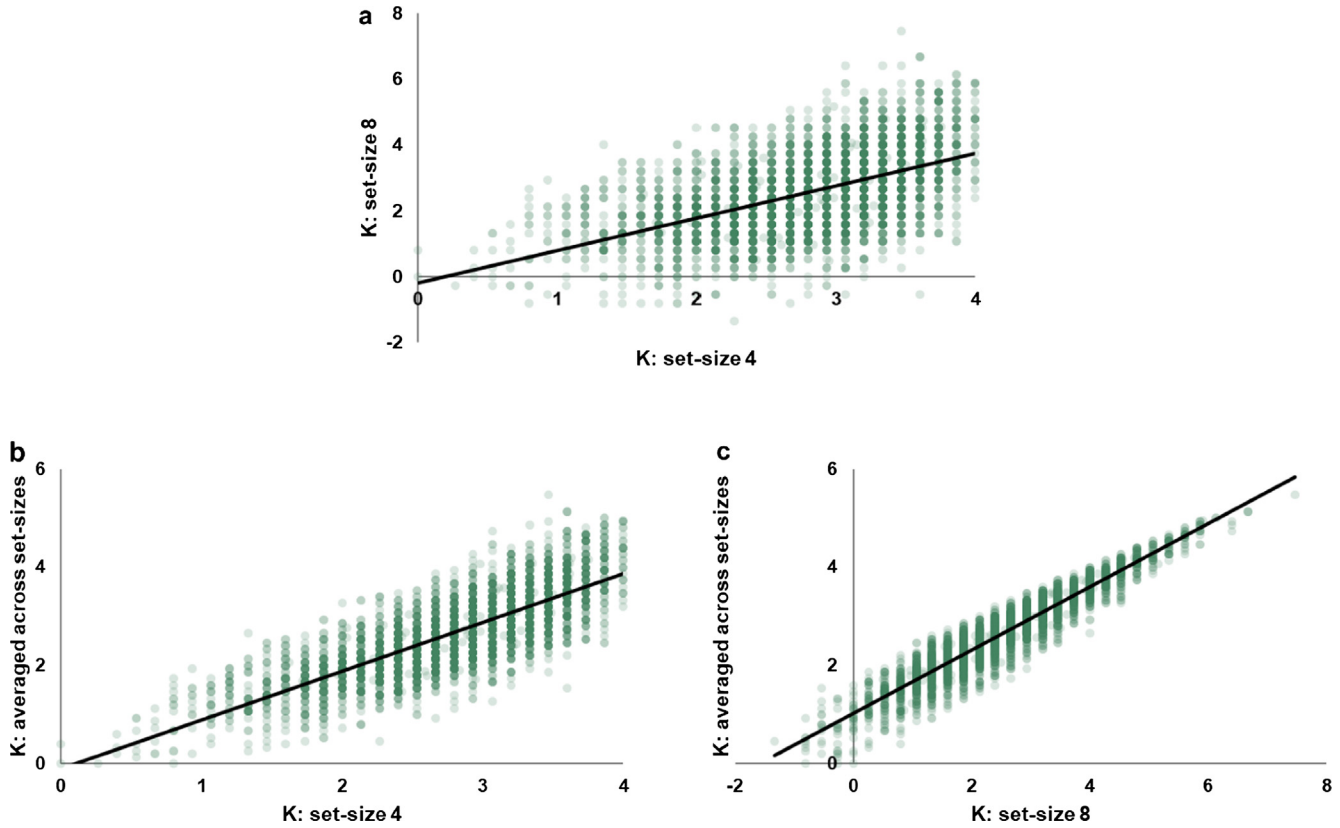

Fig. 4. The correlations between capacity estimates (K) from (a) set-sizes 4 and 8, r = 0.53; (b) set-size 4 and the average of both set-sizes, r = 0.79; (c) set-size 8 and the average of both set-sizes, r = 0.94. To better illustrate the distribution, the dots are semi-transparent, hence the darkness of an area shows the frequency of this combination of values.

similar way, thus inflating the observed correlations. To test this, we divided each data-set into two parts, each constructed of only odd- or even-numbered trials. We then calculated the correlations between the different measures from these mutually-exclusive subsets of trials, and averaged across the two correlations (e.g., the correlation between K at set-size 4 and averaged K was calculated as the average of the correlation between K at set-size 4 in odd trials and average K in even trials, and the correlation between K at size 4 in even trials and average K in odd trials). Since this results in a very small number of trials contributing to each measure, we used the Spearman-Brown correction (Brown, 1910; Spearman, 1910) to estimate the correlations for double the number of trials.

In line with the hypothesis that the strong correlations stem at least partially from the reliance on the same trials, when different trials were used to calculate each measure, the pattern of correlations was different from what we originally observed. We found correlations of r = 0.41 between K at set-size 4 and K at set-size 8, r = 0.54 between K at set-size 4 and overall K, and r = 0.58 between K at set-size 8 and overall K, lower than the correlations from all trials together. Thus, set-size 8 no longer dominated the overall K estimates. The correlation between the 4–8 drop and K at set-size 4 remained around zero (r = -0.01), and the correlations between the 4–8 drop and K at set-size 8 or overall K were lower than those based on all trials: r = −0.37 for K at set-size 8, and r = −0.28 for overall K. Critically, however, these results are difficult to interpret, because the reliability of our measures was quite low, even after the Spearman-Brown Correction was applied (r = 0.25 for the 4–8 drop, r = 0.51 for K at set-size 4, r = 0.53 for K at set-size 8, and r = 0.64 for overall K). This is expected based on previous reports that using a small number of trials will lead to a low reliability for K estimates, whether overall or by set-size, even using the Spearman-Brown correction (Xu et al., 2018), a finding which remains stable regardless of the number of participants.

Thus, our analysis raised the possibilities that capacity estimates mostly reflect the larger set-size, and that the 4–8 drop is not independent from the overall capacity estimates and from set-size 8, but more research is needed for a stronger conclusion to be drawn. If more direct evidence will support our observations, it will be in line with a view of capacity that diverges from its interpretation as simply the size of VWM-storage. Instead, as is suggested by our finding that capacity estimates are lower in the larger set-size (despite both set-sizes being above average capacity), we argue that capacity should be regarded as the ability to efficiently use VWM (Mance & Vogel, 2013). In this view, individual differences in capacity do not solely manifest a different storage-size, but instead largely reflect the differential capability to flexibly and adaptably use a similarly-limited storage.

We next turned to compare trials that included a change in the probed color to trials in which the color was the same as in the memory array. Accuracy in change trials (i.e., hit rate) was much higher than in no-change trials (i.e., correct rejection): 0.90 compared with only 0.61. This pattern was observed both for set-size 4 (0.93 vs. 0.77, respectively) and for set-size 8 (0.86 vs. 0.46, respectively). This suggests a shift of criterion, and indeed we found that the criterion was −0.41 for set-size 4 trials, and −0.69 for set-size 8 trials. This indicates that subjects had a bias to respond “different”, and this tendency was more evident in the larger set-size.

The shift in criterion is an interesting finding, which can be naturally accounted for in a slot-like view of VWM (Zhang & Luck, 2008), which posits that only a handful of items are represented with high resolution, and other items simply remain outside of VWM (it is also possible that these items are represented with extremely low resolution, as was previously suggested (van den Berg, Shin, Chou, George, & Ma, 2012), but for conciseness we focus on the simpler zero-resolution framework and return to the more elaborated alternative in the Discussion). If some items are not represented, they should be perceived as

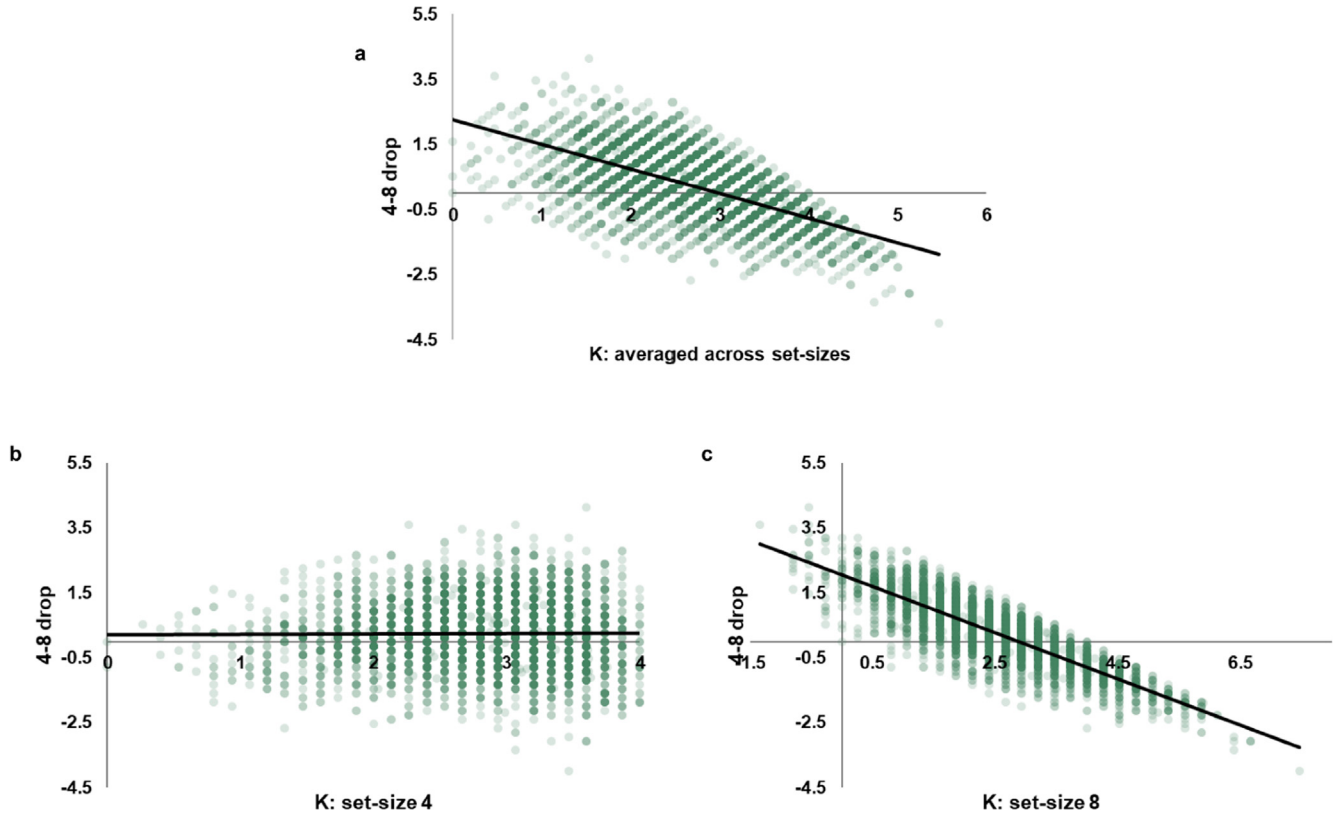

Fig. 5. The correlations between the 4–8 drop (the difference in K between set-size 4 and set-size 8; Y axis) and the three K measures (X axis): (a) K averaged across both set-sizes, r = −0.61; (b) K at set-size 4 trials, r = 0.01, (c) K at set-size 8 trials, r = −0.84. To better illustrate the distribution, the dots are semi-transparent, hence the darkness of an area shows the frequency of this combination of values.

different from the current contents of VWM even when they are unchanged, resulting in a greater tendency to respond "different", i.e., causing a shift of criterion in the observed direction. The slot-view interpretation of our observed average K is that subjects only hold an average of 2–3 items of the 4 or 8 presented colors, while all other items are completely unrepresented. Therefore, when the test array happens to present one of the unrepresented items in its original (i.e., unchanged) color, subjective experience should be that this item is *different* from those held in memory, leading to more "different" responses than there should be, i.e., the observed response bias. A slot view also predicts that the bias should be larger when more items are presented, because there are more un-stored items that subjectively resemble new items once presented in the test array.

The alternative view of VWM is of a flexible resource, such that all items enter VWM, and as more items are maintained, the resolution of the representations deteriorates (Bays & Husain, 2008). The greater representational noise that is associated with an increased set-size in this account can result in several possibilities with regards to the response bias. While accuracy should decrease as set-size increases, the predictions for the criterion are less clear. As set-size increases the representations become noisier, which might lead participants to perceive unchanged items as different from those in memory, shifting the criterion in the observed direction. However, noisier representations might necessitate stronger evidence to report a change, exactly because the presented color is always different from the perceived one, shifting the criterion in a positive direction, the opposite from what was observed here. Thus, a continuous-resource account of VWM does not necessarily predict any criterion shift, but it can accommodate it. The observed direction of the criterion shift suggests that the factors that cause a negative criterion shift are stronger than those causing a positive criterion shift, an interesting direction that can be empirically tested in future studies.

Finally, another possibility is that the observed shift of the criterion is the result of factors external to the VWM capacity debate, such as attentional load or task difficulty. For example, subjects might utilize a different strategy in the larger set-size, causing them to shift the criterion. However, this notion does not explain why the criterion was shifted to begin with (in set-size 4), or why the shift when the set-size increased was in the observed direction. Therefore, more work, perhaps involving formal model comparison, is needed to better understand the origin of the observed pattern of results. We conclude that the shifted criterion is predicted by a discrete-slot view of VWM but does not rule out a flexible-resource view of VWM and might also be the result of factors orthogonal to the structure of VWM.

To complement this signal detection analysis, we also report d', i.e., memory signal strength: d′ was 2.48 for set-size 4, and 1.14 for set-size 8, which is lower as expected. Notably, the exact amount of decrease in d′ from set-size 4 to set-size 8 might shed light on the source of this decrease. Specifically, according to a simple version of the flexible-resource view of VWM, all of the presented items are maintained in VWM, because capacity is distributed across all items. Consequently, as set-size increases, the resolution of each representation should decrease. In other words, capacity remains constant, and is divided across all N items. Therefore, on average d′ at set-size 8 should be half of d′ at set-size 4, because there are twice as many items that share the same overall capacity. Here, we found that d′ at set-size 4 was 2.48, meaning that according to the flexible resource view d′ at set-size 8 should be 1.24, which is quite close to the observed d′ of 1.14. In fact, for each participant d′ at set-size 8 should be half of d′ at set-size 4, and we can test the correlation between the actual d′ at set-size 8 and the d′ predicted based on d′ at set-size 4 (note that the predicted d′ at set-size 8 is a simple linear transformation of d′ at set-size 4), which we found to be r = 0.47 (95% CI: [0.44, 0.49]). Finding this correlation suggests that a simple version of the flexible-resource model gives a good prediction of

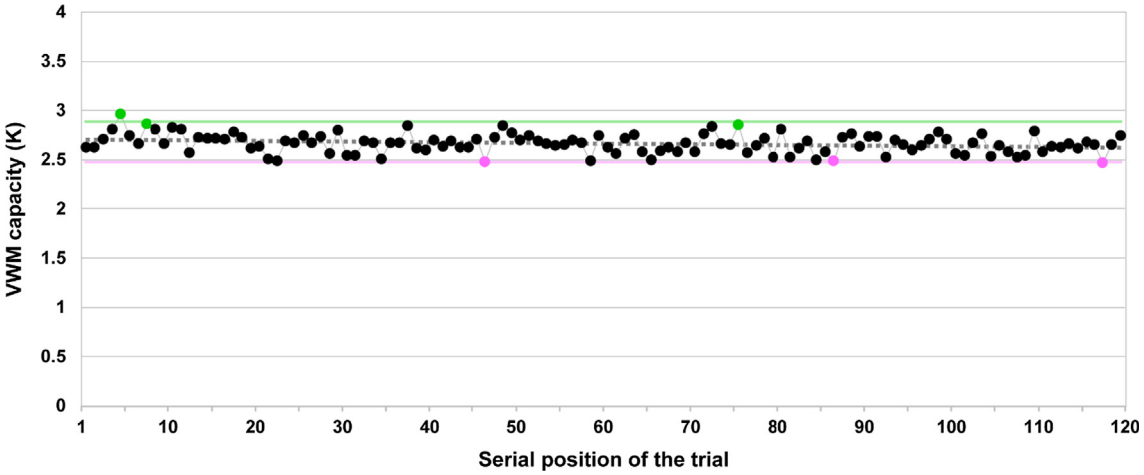

Fig. 6. K estimates, for all 3,849 data-sets pooled together, by serial position within the task. Green dots are the 3 highest K trials, with the green line showing their average. Pink dots are the 3 lowest K trials, with the pink line showing their average. The dashed grey line shows the linear relationship between K and serial position. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the decrease in d′. Versions of the flexible-resource model that incorporate other factors, such as a variable resolution, could probably produce better predictions, and in fact, in the present setting (with only two set-sizes and no additional manipulations), some sort of a flexible-resource view could have predicted any amount of a decrease in d'.

Interestingly, the alternative view, of VWM as a slot-like storage, makes very strong predictions regarding the exact level of d′ across set-sizes as well. According to this view, at each set-size some of the items are fully represented (i.e., with a very high resolution), and the rest of the items remain completely outside of VWM. This means that the observed d' at each set-size is a mixture of two d' measures: one for the fully-remembered items, and a d' of 0 for the items that were not represented. Furthermore, the slot model assumes that the d′ for the represented items should be identical between set-size 4 and set-size 8, due to the all-or-none nature of the slots.

To test this prediction, we can extract the individual d' for fully remembered items from the observed d' at set-size 4: the observed d' (2.48) times 4, divided by K at set-size 4, which gives d' = 3.56. This means that according to the slot model, the observed d′ at set-size 4 (2.48) is an average of the 2.79 items that were represented in VWM with a d′ of 3.56 and 1.21 items with d′ of 0. Because the slot model assumes that the individual-item d′ for the represented items should be identical in set-size 8, we can calculate the predicted average d' at set-size 8, because the slot model predicts that this d′ is an average of K items with d′ = 3.56 and 8-K items with d′ = 0. Thus, according to the slow model, the average d′ should be 3.56 times 2.54 (K at set-size 8), divided by 8. This gives a predicted d' at set-size 8 of 1.13, which is remarkably close to the observed d' at set-size 8, which was 1.14. This suggests that on average, the pattern of d′ across set-sizes is very well predicted by the slot model, which assumes all-or-none fixed representations.

While these results strongly support the slot model at the average level, if indeed a slot-like view of VWM can account for the pattern of d′ across set-sizes, this should be true also at the individual level. Thus, an even stronger test for the predictions of the slot model would be applying the same calculation for each data set separately and examining how well the predicted d′ correlates with the observed d′. Corroborating and extending the group-level finding, we found a very strong correlation of r = 0.83 (95% CI: [0.82, 0.84]) between the d' at set-size 8 that was actually observed and the d' predicted by the slot model. Notably, this correlation is not merely a byproduct of a tight relationship between the observed d' at set-size 4 (which was used to calculate the predicted d' at set-size 8) and the observed d' at set-size 8, because their correlation was much lower, r = 0.47 (95% CI: [0.44, 0.49]), as was found for the simple flexible-resource prediction of d′.

Thus, we found that the pattern of d′, a signal detection theory measure of memory strength, was well accounted for in a resource-like view of VWM, but it followed even more closely the predictions of a slot-like VWM which holds a limited set of items in high resolution, while other presented items remain outside of the storage. Moreover, it supports the slot model's assumption that the same memory strength is used for each represented item, regardless of the set-size. Indeed, the average d′ across set-sizes followed what is expected from the capacity estimates of K, and this was true even at the individual level. These somewhat surprising findings give novel support for the notion of a discrete, instead of resource-like, capacity limit of VWM. However, it is important to note that this analysis relies on mixing K and d′, two measures that have very different underlying assumptions (high-threshold versus signal detection). Indeed, the slot model assumes that some items will produce maximal d′ and others will have d′ = 0, which is different from the usual analysis of d′. Moreover, the current slot-model predictions of d' at set-size 8 were based on K at set-size 8, which involve the same trials. If instead we use K at set-size 4 to predict d' at set-size 8, we go back to a correlation of r = 0.47 (95% CI: [0.44, 0.49], exactly as strong as using d' at set-size 4 to predict d' at set-size 8. It is noteworthy that measures that are not based on the high-threshold assumptions of K, namely d′ and criterion, follow closely the predictions of the slot model, but more work is needed to clarify this point, and using formal model comparison could allow a future definite conclusion.

Finally, we found a correlation of r = −0.31 (95% CI: [−0.34, −0.29] between overall response bias and d′ (across the different set-sizes).

3.2. Group-level capacity estimates

We next turned to examine the dynamics of capacity throughout the span of the experiment. Comparing K values from the different serial positions of the trials (1–120) can uncover the influence of either practice or proactive interference on capacity estimates: practice should manifest in larger K values as the experiment progresses, while proactive interference will produce smaller K values in later trials. Notably, the ability to overcome proactive interference has been argued to underly at least some part of individual differences in VWM capacity, e.g., because of the need to suppress previously active items (Hartshorne, 2008; Unsworth & Engle, 2007). K estimates for each of the serial positions in the task are presented in Fig. 6.

Contrary to the interpretation of capacity as stemming from

proactive interference, we found that K estimates were extremely stable throughout the experiment, with an SD of only 0.10 between the different serial positions. As can be seen from Fig. 6, the highest K trials and lowest K trials were relatively distributed across the entire experiment, and the difference between them was not very large (mean of the 3 best serial positions: 2.89, worst: 2.47). Additionally, we found that K on the first trial of the task, where proactive interference is minimal, was already quite low (2.63), and was even slightly lower than K on the last trial, where proactive interference is maximal (2.74). To quantify the contribution of proactive interference to capacity, we calculated the Spearman correlation between serial position and K, which was negative but relatively weak, $r = -0.23$ (95% CI: $[-0.39, -0.05]$). This suggests that while some proactive interference might occur, it accounts for only 5% of the variance in capacity, in line with previous claims that it is fairly negligible in the canonical change detection task (Lin & Luck, 2012).

We found no clear evidence of improvement beyond the first ~5 trials, suggesting that after some minimal practice took place, capacity is very stable. This improvement is likely due to a slightly smaller bias (criterion was −0.55 in the first trial, and −0.54 in the fifth trial), and after this adjustment of criterion participants did not continue to improve, suggesting practice does not play a large role in capacity estimate in this short task.

As mentioned above, since VWM capacity is limited, a key question is what happens in the face of many to-be-encoded items. A simple flexible resource model should predict that capacity is evenly distributed between all items, because they are equally important and all of them can enter VWM (though in lower resolution, Bays & Husain, 2008). Contrary to this view, most present theories of VWM posit that when facing supra-capacity arrays, participants must prioritize some items at the expense of others. The unprioritized items either won't enter VWM at all (Zhang & Luck, 2008), or will enter VWM in lower resolution (van den Berg et al., 2012). Our data-set allowed us to examine whether there is systematic prioritization based on items' locations, reflected in systematic differences in accuracy between the probed locations. As can be seen from Fig. 7, accuracy is not uniformly distributed across space, and instead a handful locations are privileged. The two top-central locations produce the best accuracy, and the locations around them follow. The top part of the screen is preferred to the bottom part, while there is no systematic difference between the left and right sides of the screen. The source of the advantage for the two top-central locations could be attributed to different stages involved in the change detection task: privileged attention during encoding or less encoding noise, more stable maintenance, or easier retrieval.

A subset of participants performed a counterbalanced version of the task. Applying the spatial analysis to this subset allowed us to disentangle the previously discussed bias for "different" responses from the effect of response side, because in the regular version of our task "different" was associated with the right-hand side. For 905 counterbalanced data-sets, accuracy by the type of trial (change vs. no-change) and the response-mapping ("different" on the right vs. left) is shown in Fig. 8 (note that this analysis is much noisier than the previous one, including only about 6% of the amount of data). The results, presented in Fig. 8, suggest no clear effect of response-mapping, and importantly, both groups have a much higher hit rate than a correct rejection rate. This suggests that the response bias doesn't stem from a right-hand preference, and instead reflects a genuine shift of criterion, in line with an item-limit on VWM (see above).

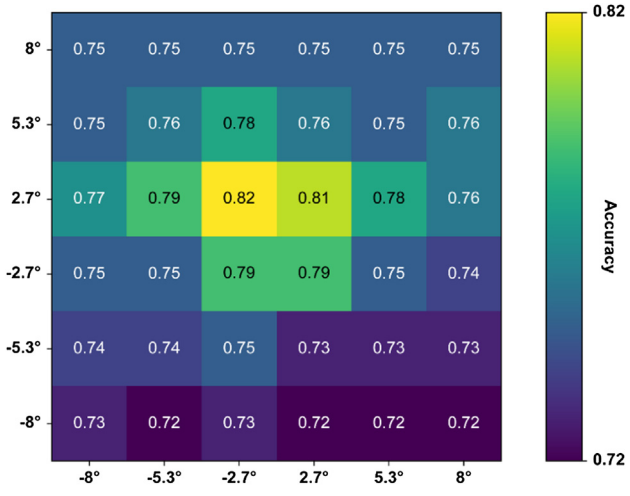

Fig. 7. A heat map of accuracy, for all 3,849 data-sets pooled together, by spatial location of the probed item.

3.3. Large-scale analysis of capacity

Our final set of analyses examined the relationship between K estimates and several factors, as a characterization of the "population-level" parameters of VWM capacity. We first examined capacity by the hour of day and month of year in which the experiment took place, and then by demographic parameters: gender, age, and field of study (since the data was not collected for the current analysis but for a range of different experiments conducted through the course of several years, the number of data-sets for which a specific measure was available differed between the reported factors).

Our data was collected throughout the year, enabling us to examine whether capacity is, for example, higher at the beginning of a semester (November and March for the winter and spring semester, respectively) than at the end of a semester (January and June), during the exams period (February and July-August), or during summer break (September-October). Similarly, we could examine whether capacity drops or rises throughout the day. As can be seen in Fig. 9, K estimates were quite stable across the year, and throughout the day, although some variation exists. The highest K month was January (the end of winter semester; mean: 2.75, SD: 0.87), and the lowest K month was June (the end of spring semester; mean: 2.51; SD: 0.85), but the range was very narrow, and the different months, both during the school year and during vacations, produced highly similar capacity estimates (SD of the different months: 0.07). As for the time of day (for this analysis we focused on 8am-5 pm, because earlier and later hours had less than 30 data-sets; this resulted in a total of 3,813 data-sets for this analysis), we found that K was slightly higher in the morning than in the afternoon, although it had a relatively restricted range. The highest K hour was 9 am (mean: 2.76, SD: 0.82) and the lowest K hour was 5 pm (mean: 2.47, SD: 0.81), with a very small range (SD of the different hours: 0.07).

As another way to characterize VWM capacity, we examined the relationship between capacity and demographic factors, starting with gender (N = 2,079, 1,456 females, 623 males). We found that females' VWM capacity was slightly higher than that of males (females' mean: 2.75, SD: 0.81; males' mean: 2.64, SD: 0.82; $BF_{10} = 2.79$). To the best of our knowledge, this interesting, though delicate, finding has not been previously reported.

Next, we examined the correlation between capacity and age (N = 1,833), because it has been shown that VWM deteriorates in old age (e.g., Jost et al., 2011). However, in our data-set of healthy young adults, with a restricted range of ages (18–39, mean 23.75), we found absolutely no relationship between capacity and age ($r = 0.01$, 95% CI: $[-0.03, 0.06]$). This shows another aspect of stability in capacity estimates using the canonical change detection task.

Finally, we examined the distribution of capacity across the different fields of study among our student participants. We divided participants into 6 groups: arts and humanities (N = 132), education and therapeutic professions (N = 150), exact sciences and engineering (N = 227), law and management (N = 150), life sciences and medicine (N = 222), and social sciences (including psychology; N = 703). We found the lowest capacity among law and management students (mean:

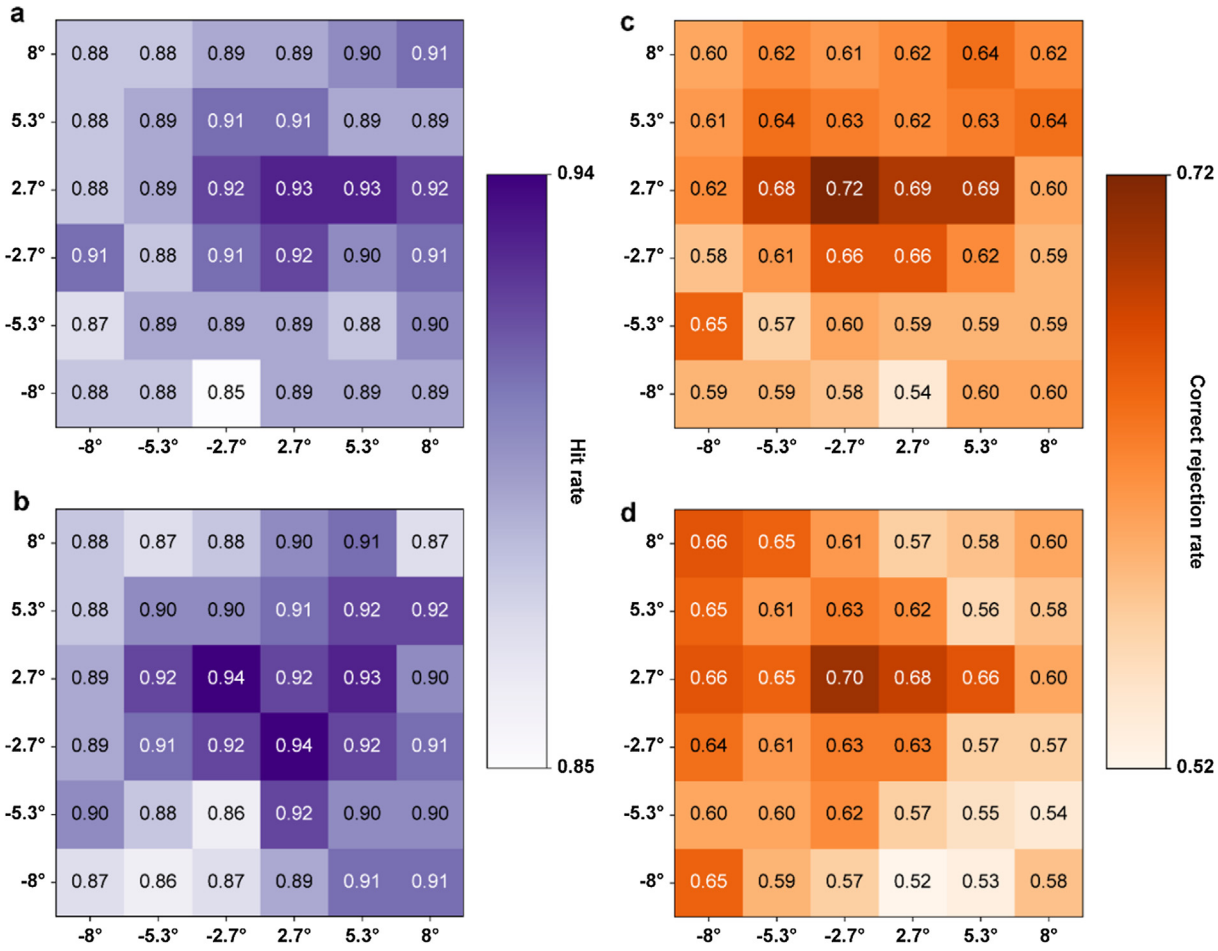

Fig. 8. A heat map of accuracy, for a subset of 905 data-sets with counterbalancing, by spatial location of the probed item, broken down by the type of trial (change vs. no-change) and the response-mapping. The left two panels depict change trials, with the "different" key on the (a) right-hand side or (b) left-hand side. The two right panels depict no-change trials, with the "different" key on the (c) left-hand side or (d) right-hand side. Note that different color-scales were used for change and no-change trials, because of the large difference in accuracy between them.

2.61, SD: 0.89, 95% CI: [2.46, 2.75]) and arts and humanities students (mean: 2.62, SD: 0.85, 95% CI: [2.48, 2.77]), medium among education and therapeutic professions students (mean: 2.69, SD: 0.84, 95% CI: [2.56, 2.83]) and exact sciences and engineering students (mean: 2.71, SD: 0.80, 95% CI: [2.61, 2.82]), and highest among social sciences students (mean: 2.75, SD: 0.81, 95% CI: [2.69, 2.81]) and life sciences and medicine students (mean: 2.84, SD: 0.86, 95% CI: [2.73, 2.95]). Thus, there are some differences among the different fields of study, but overall K estimates are relatively similar for all fields, at least for university students.

4. Discussion

By analyzing over 460,000 trials of the canonical change detection task, the present study had two goals: first, exploring the nature of VWM capacity, and second, better characterizing the change detection task itself. Our unique data-set draws a "population-level" picture of capacity limitations, which we describe below. By making our data publicly available, we hope other researchers will be able to analyze it in other ways, reaching new interesting conclusions regarding VWM.

The extremely large sample-size allowed us to uncover the fact that

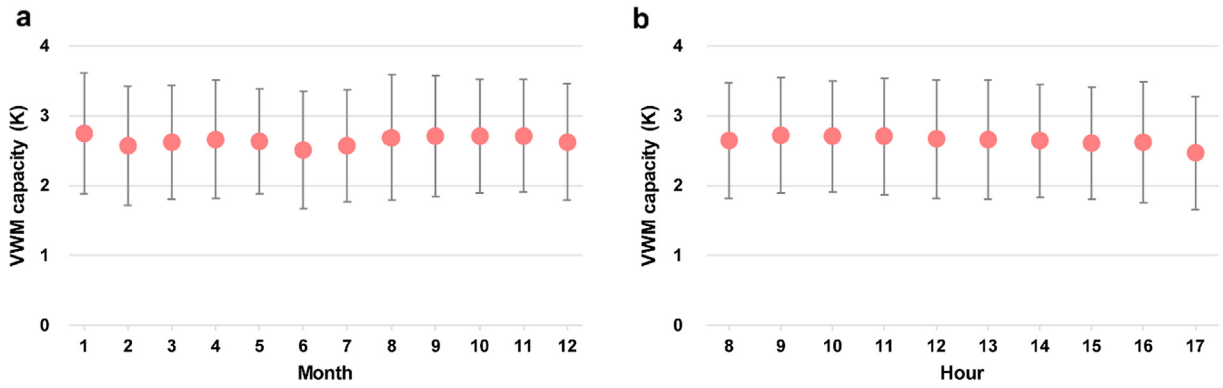

Fig. 9. Mean K estimates by (a) the month of year, and (b) the time of day in which the experiment took place. Error bars depict standard deviation.

individual K values are naturally normally distributed, which complements previous reports of tight connections between VWM capacity and factors such as fluid intelligence (Fukuda et al., 2010). Because capacity is easily quantified in a stable and meaningful way using the change detection task in its present form, we believe that measuring capacity for each participant should be the standard practice in VWM research (Luck & Vogel, 2013). We believe this will produce new insights on the sources of capacity limits, as those found in the present investigation.

Several findings support recent claims that capacity does not simply reflect the size of storage space (Fukuda et al., 2015; Mance & Vogel, 2013) and that a fundamental attribute of VWM is its active nature. K estimates were lower for set-size 4 than for set-size 8, suggesting that not only storage-size but other factors (e.g., selecting which items to encode) influence VWM processes. We found initial evidence that capacity estimates are affected by these factors, as manifested in the stronger effect set-size 8 trials have on overall K than set-size 4 trials. Furthermore, the 4–8 drop, previously used to asses attentional control independently from storage-size, significantly correlated with K, suggesting that similar processes dominate the two measures. While the findings based on correlations need more direct evidence to be considered convincing, they are in line with the capacity reduction in the larger set-size, both suggesting that there's more to capacity than mere storage-size. We therefore argue that a central cause for individual differences in VWM capacity is the ability to flexibly and adaptably use the VWM storage space, that might be similar in size among individuals. It could be that most individuals can store about 3 simple items, but low-K individuals store only 1 task-relevant item and 2 completely irrelevant items, while high-K individuals efficiently select 3 task-relevant items, and perhaps even more by chunking several items together or by making better use of ensemble representations (Brady & Alvarez, 2011; Haberman, Brady, & Alvarez, 2015). These abilities would become critical in complex real-life situations, which include vast amounts of meaningful information from which only a small subset should be carefully selected to enter VWM. This likely contributes to the tight connections between VWM and attentional control in a range of situations (Vogel et al., 2005).

Despite the central role attentional control plays in individual differences of VWM capacity, it is not likely the source of capacity limit itself. Namely, we found that proactive interference, i.e., the ability to prevent previously relevant items from disrupting the maintenance of currently relevant items, doesn't explain the low average capacity, unlike previous claims (Unsworth & Engle, 2007). The correlation between K and the serial position within the task was quite low (accounting for ~5% of K variability across the task), and K was low already in the first trial (in fact, slightly lower than in the last trial). Thus, VWM is genuinely limited in capacity, even without proactive interference (Lin & Luck, 2012).

A central issue in VWM research is how to best describe its capacity limits. One leading approach views capacity as a continuous resource that can be shared among an arbitrary number of representations (Ma et al., 2014), trading quantity for quality as more items are added or as their complexity increases (Alvarez & Cavanagh, 2004). The contrasting view is that VWM has a discrete set of place-holders that can each maintain one item (regardless of its complexity), and once all slots are allocated additional items are left completely outside of VWM (Vogel & Machizawa, 2004; Zhang & Luck, 2008). This is admittedly a simplified description of a complicated debate, and many subtle intermediate positions have been proposed (Brady et al., 2011). Regardless, the question of whether it is fruitful to conceptualize VWM as holding a very small set of items remains an important one. Our results confirmed one prediction of a strict item limit, in the form of a criterion bias towards "different" responses. If some of the items simply do not enter VWM, it is natural that subjects will treat them as not matching the contents of VWM when they are presented again (unchanged) at test. Accordingly, the bias was larger for 8 items than for 4, meaning when more items are presumably left outside of VWM. While it is possible that the additional items were held in VWM with extremely low resolution (van den Berg et al., 2012), if their representations are so corrupted they are unusable even when changes are very large as in the present paradigm, it might be sufficient to treat them as effectively absent from VWM under circumstances similar to the change detection task employed here. It may be that in the real world, regardless of whether VWM operates in a slot-like or resource-like manner, observers act as if the representations are discrete given specific constraints. In further support for a slot-like view of VWM, such that some items are represented with very high resolution while the others are left completely outside of VWM, we found a decrease in d' (the signal detection theory measure of memory signal strength) with increased set-size in a manner compatible with the predictions drawn from this discrete capacity assumption. Namely, d' at set-size 8 could be well predicted, at both the average and the individual levels, from the observed d' at set-size 4 and the K estimates at both set-sizes. This suggests that d′ at each set-size N reflects a mixture of K items that have high resolution that is fixed across set-sizes, and N-K items that are not represented at all in VWM. Notably, this analysis was based on mixing two measures with very different underlying assumptions: K, which is a high-threshold measure, and d′, which is a signal detection measure. Future formal model comparisons could help clarify the source of the decrease in d′.

Importantly, while the pattern of the signal detection measures, i.e., the criterion shift and d′, is naturally predicted by a slot-like view of VWM, it can also be accounted for in a flexible-resource view, and might even be the result of factors external to the structure of VWM. Thus, more research is needed to explore the criterion and d′ pattern in the classic change detection task, but we do note that these measures, which are not based on high-threshold assumptions, seem to offer initial novel support for the predictions of a discrete slot-like structure of VWM.

The present capacity estimations agree with numerous findings (Cowan, 2001) that place the average capacity limit at around 3 simple items, with vast individual differences (Luck & Vogel, 2013). Our extremely large sample-size allowed us to uncover two novel characteristics of K estimates at the "population-level": the normal shape of the distribution of individual K estimates (see above), and the fact that females have a slightly higher capacity than males.

Aside from new insights on VWM capacity, our data-set allowed us to better characterize the canonical change detection task itself, highlighting important issues that should be kept in mind when using the paradigm to investigate capacity. First, it is critical to include set-sizes that are well above average capacity limits, as we found that K is dominated by estimates from set-size 8 trials, although using only very large set-sizes is not recommended because of the expected motivational effects. Second, to remove the effect of practice, the task should include about 10–12 practice trials to allow participants to adjust their criterion (we had ~6 practice trials and found improvement for another 5 trials). After this phase, practice has no impact, at least when the task is as short as here. Third, our heat map revealed several locations that benefit from greater attentional allocation, specifically the ones closest to fixation on the top of the screen, followed by the locations surrounding them. Performance was comparable at the right and left side, and better above than below fixation. These preferences might arise at one of several possible processing stages, and perhaps at multiple stages together. The privileged items might have more attention or less noise at encoding, better maintenance once in VWM, or easier retrieval, and these options are not mutually exclusive. The source of the spatial preference we found could be the target of future research. Finally, counterbalancing the response keys doesn't seem to be necessary.

Importantly, all our data and conclusions are relevant to the classic same-different change detection paradigm. In a continuous-report recall version of this task (Wilken & Ma, 2004; Zhang & Luck, 2008), one can disentangle the probability that an item enters VWM from the resolution with which it is represented. It is not yet known whether this

version correlates with aptitude measure as the classical version does, and if so which of the two factors of capacity drives this correlation (though evidence from the canonical change detection task suggested that it is storage-size and not resolution that correlates with intelligence; Fukuda et al., 2010), which calls for future research. It is perfectly plausible that analyzing the recall change detection task will reveal support for a more resource-like view of VWM. Notably, K from change detection performance is tightly correlated with the probability of remembering an item derived from continuous report (Zhang & Luck, 2008), which is another way to quantify the number of items held in VWM. This suggests that at least some of the conclusions drawn on K will be generalized to other methods and ways of measuring capacity.

To conclude, the change detection task is an excellent way to measure capacity, being stable not only within an individual, but also across the population: K was highly similar between different months of the year, hours of the day, and, in our sample of normal students, even ages. While it is critical to understand the task's limitations and not confuse it with the theoretical construct of VWM capacity, within these limitations (e.g., paying attention to the use of complex items; Awh et al., 2007, see the Introduction) the now-classic change detection paradigm indeed deserves its canonical status.

Declaration of Competing Interest

None.

Acknowledgments

This work was supported by the Israel Science Foundation (grant number 862/17 awarded to Roy Luria) and the Azrieli Fellowship (awarded to Halely Balaban). We wish to thank all past and present lab members, especially Ayala Allon, Maya Ankaoua, Hagar Cohen, Britt Hadar, Shiri Tascesme, and Anna Vaskevich, for data collection.

References

Allon, A. S., & Luria, R. (2017). Compensation mechanisms that improve distractor filtering are short-lived. *Cognition, 164*, 74–86.

Allon, A. S., Vixman, G., & Luria, R. (2018). Gestalt grouping cues can improve filtering performance in visual working memory. *Psychological Research,* 1–17.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15*(2), 106–111 01502006 [pii].

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 PSCI1949 [pii].

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321*(5890), 851–854. https://doi.org/10.1126/science.1158023.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science, 22*(3), 384–392.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision, 11*(5), 4. https://doi.org/10.1167/11.5.4.

Brown, W. (1910). Some experimental results in the correlation of mental abilities 1. *British Journal of Psychology, 3*(3), 296–322.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences, 24*(1), 87–114 discussion 114–85.

Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51*(1), 42–100 S0010-0285(05)00002-2 [pii].

D'agostino, R. B., Belanger, A., & D'agostino, R. B., Jr (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician, 44*(4), 316–321.

Fukuda, K., Woodman, G. F., & Vogel, E. K. (2015). Individual differences in visual working memory capacity: Contributions of attentional control to storage. *Mechanisms of Sensory Working Memory: Attention and Perfomance, XXV*, 105.

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. https://doi.org/10.3758/17.5.673.

Haberman, J., Brady, T. F., & Alvarez, G. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General, 144*(2), 432.

Hartshorne, J. K. (2008). Visual working memory capacity and proactive interference. *PLoS One, 3*(7), e2716.

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*(2), 220–229. https://doi.org/10.1037/a0032060.

Jost, K., Bryck, R. L., Vogel, E. K., & Mayr, U. (2011). Are old adults just like low working memory young adults? Filtering efficiency and age differences in visual working memory. *Cerebral Cortex, 21*(5), 1147–1154. https://doi.org/10.1093/cercor/bhq185.

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

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279–281. https://doi.org/10.1038/36846.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences, 17*(8), 391–400. https://doi.org/10.1016/j.tics.2013.06.006.

Keshvari, S., van den Berg, R., & Ma, W. J. (2013). No evidence for an item limit in change detection. *PLoS Computational Biology, 9*(2), e100297.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17*(3), 347–356. https://doi.org/10.1038/nn.3655.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's.* Guide Psychology Press.

Mance, I., & Vogel, E. K. (2013). Visual working memory. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(2), 179–190.

Martinussen, R., Hayden, J., Hogg-Johnson, S., & Tannock, R. (2005). A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 44*(4), 377–384 S0890-8567(09)61489-1 [pii].

Parra, M. A., Sala, S. D., Abrahams, S., Logie, R. H., Mendez, L. G., & Lopera, F. (2011). Specific deficit of colour-colour short-term memory binding in sporadic and familial alzheimer's disease. *Neuropsychologia, 49*(7), 1943–1952. https://doi.org/10.1016/j.neuropsychologia.2011.03.022.

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

Phillips, W. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics, 16*(2), 283–290.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271–295.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*(1), 104.

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2015). Working memory delay activity predicts individual differences in cognitive abilities. *Journal of Cognitive Neuroscience, 27*(5), 853–865.

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.

van den Berg, R., Shin, H., Chou, W. C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences of the United States of America, 109*(22), 8780–8785. https://doi.org/10.1073/pnas.1117465109.

Vaskevich, A., & Luria, R. (2018). Adding statistical regularity results in a global slowdown in visual search. *Cognition, 174*, 19–27.

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature, 428*(6984), 748–751. https://doi.org/10.1038/nature02447.

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 nature04171 [pii].

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 27*(1), 92–114.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12) 11 11.

Xu, Z., Adam, K., Fang, X., & Vogel, E. (2018). The reliability and stability of visual working memory capacity. *Behavior Research Methods, 50*(2), 576–588.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233–235. https://doi.org/10.1038/nature06860.