Research Article

Psychological Science
1–15
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797618813538
www.psychologicalscience.org/PS

Working Memory Has Better Fidelity Than Long-Term Memory: The Fidelity Constraint Is Not a General Property of Memory After All

Natalie Biderman[1], Roy Luria[1,2], Andrei R. Teodorescu[3], Ron Hajaj[1], and Yonatan Goshen-Gottstein[1]
[1]School of Psychological Sciences, Tel Aviv University; [2]Sagol School of Neuroscience, Tel Aviv University; and [3]Department of Psychology, University of Haifa

Abstract

How detailed are long-term-memory representations compared with working memory representations? Recent research has found an equal fidelity bound for both memory systems, suggesting a novel general constraint on memory. Here, we assessed the replicability of this discovery. Participants (total $N = 72$) were presented with colored real-life objects and were asked to recall the colors using a continuous color wheel. Deviations from study colors were modeled to generate two estimates of color memory: the variability of remembered colors—fidelity—and the probability of forgetting the color. Estimating model parameters using both maximum-likelihood estimation and Bayesian hierarchical modeling, we found that working memory had better fidelity than long-term memory (Experiments 1 and 2). Furthermore, within each system, fidelity worsened as a function of time-correlated mechanisms (Experiments 2 and 3). We conclude that fidelity is subject to decline across and within memory systems. Thus, the justification for a general fidelity constraint in memory does not seem to be valid.

Keywords
fidelity, working memory, long-term memory, continuous-report paradigm, color memory, replication, open data, preregistered

Received 4/16/18; Revision accepted 9/7/18

An undeniable empirical fact is that memory declines with the passage of time, over spans ranging from seconds to years. Indeed, many theorists (e.g., Atkinson & Shiffrin, 1968; Cowan, 2005; Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005; Talmi, Grady, Goshen-Gottstein, & Moscovitch, 2005) argue for unique memory systems—working memory (WM) and long-term memory (LTM)—characterized, among other things, by their rate of information decline (but see, e.g., Brown, Neath, & Chater, 2007, and Howard & Kahana, 2002, for frameworks that reject the WM–LTM distinction).

Understanding the mechanisms mediating mnemonic-information decline has been the subject of heated debates. Over the years, a multitude of time-correlated mechanisms have been proposed for either or both WM and LTM, including decay (Sadeh, Ozubko, Winocur, & Moscovitch, 2016), interference (Oberauer & Lin, 2017; Underwood, 1957), distinctiveness (Brown et al., 2007), and inhibitory control (Anderson, 2003), to name but a few. All of these mechanisms are *time correlated*, in that they naturally cooccur with the passage of time in both real-life and experimental settings. Thus, mnemonic information typically declines as a function of time, but time per se is not necessarily the cause of this decline.

Commonly, memory is indexed by the probability of retrieving an event. Yet stored information comprises

Corresponding Author:
Natalie Biderman, Columbia University, Department of Psychology, 406 Schermerhorn Hall, 1190 Amsterdam Ave.,New York, NY 10027
E-mail: natalie.biderman@columbia.edu

not only remembering the mere occurrence of an event but also the fidelity by which this event is remembered (Brady, Konkle, & Alvarez, 2011). Because fidelity is a type of mnemonic information, it seems reasonable to predict that it, too, would be affected by time-correlated mechanisms. If so, fidelity should be worse for LTM, in which memories are tested over longer time spans, than for WM (e.g., Hollingworth, 2004; Schurgin & Flombaum, 2015). Differences in fidelity should likewise be observed when examined at different durations, even within WM and within LTM.

A novel approach to assessing fidelity is the *continuous-report paradigm* (e.g., Wilken & Ma, 2004; Zhang & Luck, 2008), wherein participants are asked to reproduce a particular feature (e.g., color) of previously encoded stimuli on a continuous response scale. A mixture model is then used to fit the response errors—the difference between study and report color—and to produce two estimates of color memory: (a) the probability of guessing the color—referred to as a *PG estimate*—and (b) the fidelity for remembered colors—referred to as an *SD estimate* (Zhang & Luck, 2008; cf. Bays, Catalao, & Husain, 2009). The SD estimate is inversely related to memory fidelity; specifically, the higher the SD estimate, the lower the memory fidelity.

Several WM studies have reported a dissociation between the SD and PG estimates when the number of to-be-remembered items (set size) in the study array increased (e.g., Zhang & Luck, 2008; but see Bays & Husain, 2008). Whereas guessing rates increased gradually as a function of set size, fidelity worsened over the first three items to a plateau—a lower limit on fidelity (i.e., the fidelity of three items equaled that of six; but see Van den Berg & Ma, 2014). Theoretical models have attributed the fidelity plateau to intrinsic properties of WM (e.g., number of slots, Zhang & Luck, 2008, 2009; quantity of resources, Bays & Husain, 2008).

Critically, if the lower limit on fidelity is explained in terms of WM architecture, then it should not be identical to the lower limit in LTM, in which a different architecture, operating over greater time scales, mediates performance. This was recently falsified by Brady, Konkle, Gill, Oliva, and Alvarez (2013), who compared color fidelity of real-life objects in WM with that in LTM. As expected, the color-guessing rates (PGs) were higher in LTM than in WM. Yet, strikingly, for remembered colors, the identical fidelity plateau was found in the two memory systems—specifically, the SD estimate was at approximately 20°. Moreover, this identical fidelity limit was also observed within each system (Control Experiments 1–4; Brady et al., 2013). On the basis of the equal WM–LTM fidelity limit, Brady and colleagues concluded that the lower limit on fidelity is a general property of memory.

Importantly, Brady et al.'s (2013) novel and unintuitive claim was based on confirming the null hypothesis using null-hypothesis significance testing, with an underpowered sample size (*N*s = 5 and 9 in their Experiments 1a and 1b, respectively). The current study was devised to assess whether, indeed, a single constraint underlies fidelity across and within memory systems. We compared SD estimates across WM and LTM (Experiments 1 and 2) and examined whether fidelity worsened as a function of time-correlated mechanisms separately within each system (Experiments 2 and 3).

Following Simonsohn (2015), we predetermined a sample size in all of our experiments (*N*s = 24) that was substantially larger than Brady et al.'s (2013). We included Bayesian hypothesis testing (Bayes factor, or BF) to quantify the evidence in favor of the null hypothesis or of our hypotheses. Finally, to estimate SD and PG parameters, we used both maximum-likelihood estimation (as did Brady et al.) and Bayesian hierarchical modeling. Experiment 3 was preregistered on the Open Science Framework (osf.io/ksr4d).

Experiment 1

Here, we compared the fidelity of WM and LTM using a design similar to that used by Brady et al. (2013, Experiment 1a), with the following two main exceptions. First, Brady et al. included three items in the WM study array, all of which were sequentially probed for color memory (whole-report procedure). The three reports were then aggregated to a response-error distribution from which a single SD estimate for WM was assessed and compared with the SD estimate for LTM (in which 180 items were studied and then tested). Because reporting a color may induce output interference, the fidelity of the third WM report may be worse than that of the first (Adam, Vogel, & Awh, 2017; Peters et al., 2018). Consequently, the aggregated SD estimate for WM may have erroneously appeared to be similar to the SD estimate for LTM. Furthermore, the WM fidelity limit was previously determined with only a single test probe (partial-report procedure). To get a purer measure of WM fidelity and to conform to the standard procedure, in Experiment 1, we probed only a single item in each WM trial. The SD estimate for WM we derived from our procedure could be compared with the SD estimate for LTM, given that earlier findings had demonstrated that three encoded items with a single test probe were sufficient to reach the WM fidelity limit (Zhang & Luck, 2008, 2009).

Second, Brady et al. (2013) tested color memory for studied items without verifying that participants remembered the items themselves. In contrast, we asked participants to first make judgments about whether stimuli were old or new and to subsequently retrieve their color. Color memory was estimated only for hit trials, thus disregarding color performance for items for which item-specific mnemonic information was weak or unavailable.

Method

Participants. Because the finding of equal WM–LTM fidelity by Brady et al. (2013) was essentially a null effect, we could not compute an a priori sample size on the basis of a power analysis. Nevertheless, in all of our experiments, we predefined the sample size ($N = 24$) such that it would be 4.8 and 2.7 times larger (Simonsohn, 2015) than the sample size of Brady et al.'s Experiments 1a ($N = 5$) and 1b ($N = 9$), respectively. Thus, 24 Tel Aviv University students participated in Experiment 1 for course credit (20 women; 23 right-handed; age: M = 22.33 years, SEM = 0.62). One participant was excluded from the analyses because he met the first predefined exclusion criterion (see the Analysis section). In all experiments reported here, participants had normal or corrected-to-normal vision and normal color vision (assessed using the EnChroma Color Blind Test; http://enchroma.com/test/instructions/).

Apparatus and stimuli. Stimuli were presented on an LCD monitor (24-in. LG; 1,920 × 1,080 resolution; 60-Hz refresh rate) using MATLAB (The MathWorks, Natick, MA) and the Psychophysics Toolbox Version 3 (Brainard, 1997). Participants sat approximately 60 cm away from the screen. Responses were collected via the computer keyboard.

In all, 540 images of categorically distinct objects were selected from the stimulus set used by Brady, Konkle, Alvarez, and Oliva (2008; downloaded from https://bradylab.ucsd.edu/stimuli.html). None of the objects had any discernable association with a unique color (e.g., "banana" was excluded from the object pool). All images subtended approximately 6° of visual angle.

Following the recommendation of T. Brady (personal communication, September 6, 2016), the start-off color of all images was unified to the same hue in the Commission Internationale de l'Éclairage (CIE) L*a*b* color space (see our MATLAB scripts for hue transformation at osf.io/93cvs). This standardization awarded a perceptually uniform color space throughout the experiment, which simplified interpretation of report angles and comparisons across objects. Colors were then assigned by adding a random angle between 0° and 360° to the original hue (see Section SM1 in the Supplemental Material available online).

The stimuli were divided into two experimental conditions (WM, LTM), with 270 images per condition. In each condition, 180 images were designated to be "old" items (targets) and 90 to be "new" items at test (lures). The stimuli were counterbalanced such that across participants, each image appeared an equal number of times in each experimental condition and appeared two thirds of the time at test as a target and one third as a lure. Stimuli were presented in a different random order for each participant.

Procedure and design. The design included two memory conditions, manipulated within participants: WM and LTM. The order of conditions was counterbalanced across participants. Each condition was preceded by a corresponding practice block.

WM. Each trial began with a fixation cross presented at the center of the screen for 350 ms. Next, the study display appeared for 3 s; the display consisted of three simultaneously presented colored images in a triangular formation. Participants were asked to remember both the identity of the item and its color. Subsequently, a fixation cross appeared for 1 s, followed by two test displays that remained until response (see Fig. 1). At test, only a single image was presented in gray in one of the three study locations. Participants were first asked to judge whether the image was old (right arrow) or new (left arrow). Then, they were asked to retrieve the color of the studied object using a continuous color wheel. They were told that at times people mistakenly judge old objects as new. Thus, even if they had judged the item as new, they nevertheless were to choose the color that seemed "most suitable" for that object. These instructions were employed because we initially sought to analyze color memory for false alarms and for misses (yet it turned out that we did not have enough trials to extract reliable estimates). To-be-tested items at study and lure items at test appeared equally often in each of the possible three locations. The order of target and lure test trials was randomly assigned for each participant.

Because we had a limited number of images, items had to be repeated in the WM condition. To this end, the WM session was divided into four blocks, such that items were not repeated within a block yet were repeated up to four times across blocks (for further details, see Section SM2 in the Supplemental Material).

LTM. During the study phase, participants viewed 180 images presented one at a time at the center of the screen for 3 s each. Study trials were separated by a 1-s retention interval that included a fixation cross (see Fig. 1). Participants were instructed to remember both the identity and color of all objects. To direct participants' attention to both the objects and their colors, we asked participants to classify how probable it was that they would encounter each object in the real world in the color in which it was presented. Responses were made on a 3-point scale

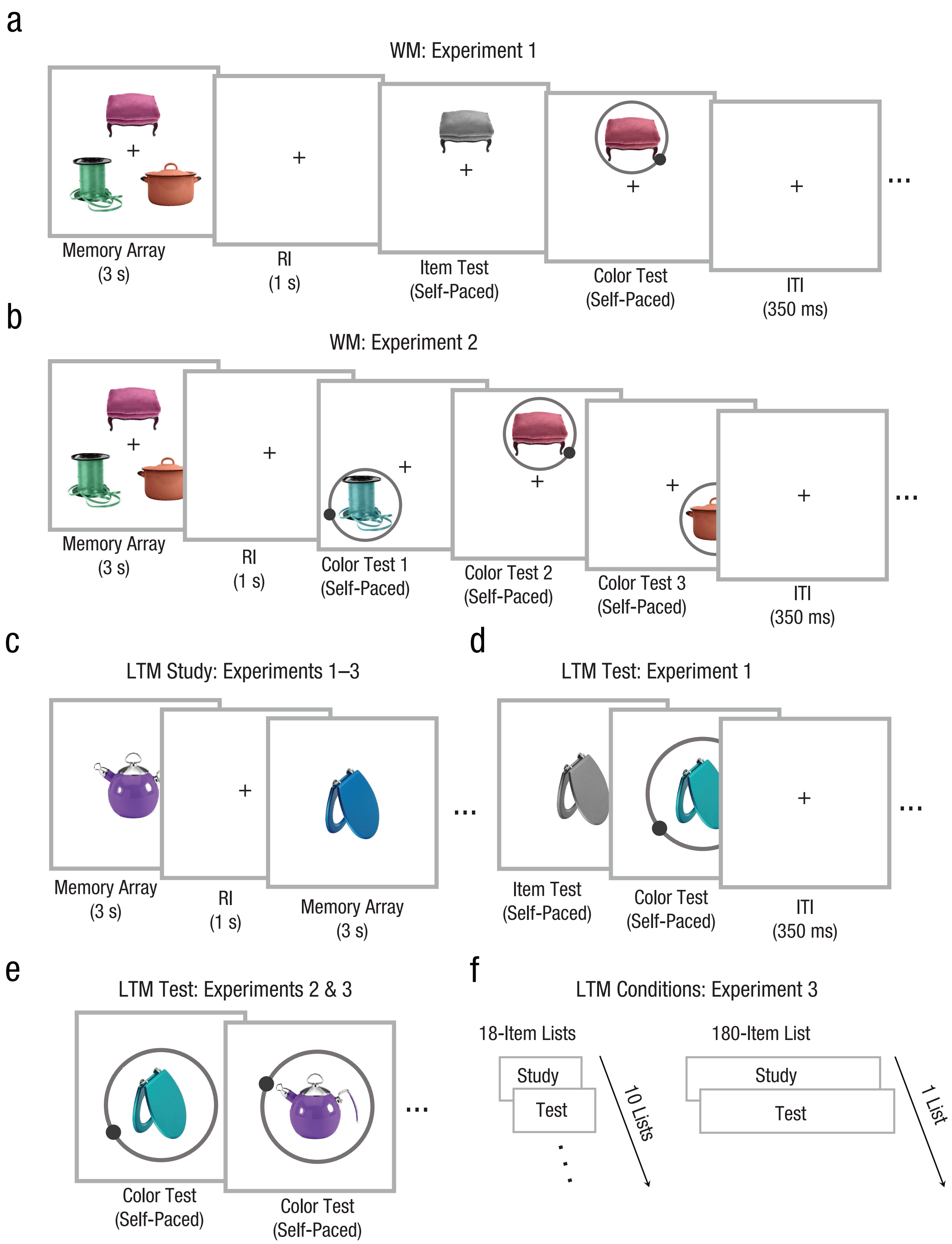

Fig. 1. Sequence of events and experimental conditions in Experiments 1 to 3. In Experiment 1, the test phase in both the working memory (WM) condition (a) and long-term memory (LTM) condition (d) included a binary (old/new) item-memory test and then a continuous color-memory test. For the WM condition, only a single item was probed. In Experiments 2 and 3 (b and e), a color-memory test alone was administered. In the WM condition of Experiment 2, each of the three studied items was successively probed (b). Experiment 3 included LTM conditions with varying memory lists (f). Each list included a study phase (c), followed by a test phase (e). RI = retention interval; ITI = intertrial interval.

(1 = *very improbable*, 2 = *pretty probable*, 3 = *very probable*) using the corresponding numerical keys on the keyboard.

The test phase immediately followed and included 180 target and 90 lure trials, presented in a random order. The test displays were identical to those in the WM condition, including the positioning of the target item at test in the identical spatial location as at study.

Continuous-report procedure for color memory. After the old/new judgment, a gray circle (radius = 4.7° of visual angle) appeared around the image with a mouse pointer situated at its center (see Fig. 1). The color of the image changed continuously as participants moved the mouse around the circle. The hue of the item was determined by the angle between 0° and an imaginary line between the center of the item and the mouse pointer position. Participants were instructed to reproduce the studied color of the item as accurately as possible by moving and clicking the mouse at their own pace. The deviations of report colors from studied colors (in Experiment 1, the reference color was the predefined color of targets and lures; see Section SM1) were defined as response errors, ranging from −180° to 180°.

Feedback was given on hit trials, with the words “good,” “great,” or “perfect” appearing on the screen for response errors of less than 10°, 5°, or 1°, respectively (cf. Brady et al., 2013). To increase participants’ motivation, we gave (false) positive feedback on false-alarm trials, with the word “good” appearing with a probability of .25.

Analysis. In Experiment 1, color memory was analyzed only on hit trials for item memory (WM hit rate: M = 96.62%, SEM = 0.72; LTM hit rate: M = 83.45%, SEM = 1.73). Importantly, when adding miss trials—the only other type of trials in which color information was registered—to the analyses, the main findings of Experiment 1 did not change for any of the color-memory estimates (all ts > 5.99, ps < .001, and BF_{10}s > 4.1×10^3). In Experiments 2 and 3, in which item memory was not tested, color memory was analyzed for all trials.

Color-memory estimations. In all experiments, two dependent measures were used to estimate color memory in each experimental condition: (a) estimates of color fidelity (SD) and (b) estimates of the probability of guessing the color (PG, with 1 − PG representing the probability of remembering the color). These estimates were computed using the two-parameter mixture model of Zhang and Luck (2008). The Zhang-Luck model assumes that participants either remember the study color with some degree of precision or do not remember it at all. Accordingly, response-error distributions (ranging from −180° to 180°) are assumed to be a mixture of two distributions: (a) a circular-normal (von Mises) distribution around the study color, with its standard deviation (SD estimate) inversely related to the precision of the representation (higher SDs correspond to worse precision), and (b) a uniform distribution, with its mixture weight reflecting the guessing-probability (PG estimate).

Notably, according to more complex models (e.g., Bays et al., 2009), the second component includes sources further to pure guessing (e.g., retrieval of distractors’ color from the study array). Here, we adopted the Zhang-Luck model because this was the model of choice adopted by Brady et al. (2013), whose findings are the focus of the current replication attempt. In addition, these alternate models were designed for the analysis of WM data that include distractors in the study array. Thus, their application to LTM data is nontrivial (but see, e.g., Richter, Cooper, Bays, & Simons, 2016).

Following the data-analysis scheme used by Brady et al. (2013), we computed the Zhang-Luck model parameters—SD and PG estimates—that maximized the likelihood of the data under the model (using the *MLE* function in MATLAB MemToolbox; Suchow, Brady, Fougnie, & Alvarez, 2013). These parameters were calculated for every participant in each experimental condition. The difference between experimental conditions was then tested using a paired-samples t test or a repeated measures analysis of variance (ANOVA). In addition, we evaluated the trustworthiness of our effects using a combination of Bayesian hypothesis testing (BFs) and effect-size estimates with confidence intervals (CIs; Lakens, 2013; detailed below). We also computed the cumulative binomial probability (CBP) of getting our prediction under the null hypothesis (e.g., in Experiment 1, the probability of getting better performance in WM than in LTM was .5).

Bayesian hypothesis testing. Frequentist null-hypothesis significance testing was complemented with Bayesian hypothesis testing, which quantified the evidence for the presence or absence of effects. We calculated BFs using the *BayesFactor* package (Version 0.9.11-1; Morey & Rouder, 2015) for the R software environment (R Core Team, 2015). For mean comparisons, we used the t-test BF function with default settings (medium prior scale). For factorial analyses, we used the ANOVA BF function with default settings (medium prior for fixed effects and nuisance prior for random effects), with the participant factor considered random in within-participants designs. A BF_{10} depicts how much more likely the data are on the assumption of an effect (i.e., H_1) compared with an assumption of the null hypothesis (i.e., H_0). A BF_{01} portrays how much more likely the data are on the assumption of the null hypothesis over the assumption of an effect. We adopted the convention that a BF between 1 and 3 depicts anecdotal evidence (or inconclusive data) for the hypothesis in question, a BF greater than 3 provides moderate evidence (i.e., one hypothesis is 3 times more likely than the other), a BF greater than 10 suggests strong evidence, and a BF greater than 100 implies extremely strong evidence (Lee & Wagenmakers, 2013).

Effect sizes and CIs. To assess effect sizes for both *t* tests and ANOVAs, we computed partial eta squared with 90% CIs (Lakens, 2013). CIs were calculated using the *MBESS* package in the R environment (Kelley, 2007), including the correction for a within-participants design (see http://daniellakens.blogspot.co.il/2014/06/calculating-confidence-intervals-for.html).

Exclusion criteria. In all experiments, we used two exclusion criteria (preregistered for Experiment 3; see osf.io/ksr4d) applied on the maximum-likelihood estimates. The first was an inaccurate fitting of the mixture model. Specifically, high guessing rates may erroneously be estimated by a very low PG estimate combined with a very wide SD estimate (Suchow et al., 2013). To discover this pattern, we first searched for a combination of an SD estimate greater than 80° and a PG estimate of less than .05 for each participant, in each experimental condition. Then, to confirm the problematic pattern for participants who met this criterion, we graphed the histogram figures and model fits to verify whether, indeed, the von Mises distribution was wide and the uniform distribution of guesses was high—a pattern that does not correspond to a low PG estimate. The data of participants who demonstrated this pattern, in any experimental condition, were excluded from the analyses. Only a single participant in Experiment 1 met this criterion.

The second exclusion criterion was subsequently applied. The data of participants with a mean performance 3 standard deviations above their group mean (representing poor color memory) for any of the two color-memory estimates in any experimental condition were also excluded. Three participants met this criterion in Experiment 2, and 1 participant met this criterion in Experiment 3.

Notably, the exclusion of participants did not change the results. In all experiments, when the entire sample size (N = 24) was included in the analyses, the main findings in each experiment remained significant for all color-memory estimates (for paired comparisons, all ts > 2.74, ps < .012, and BF_{10}s > 4.32; for three-group comparisons, all Fs > 7.07, ps < .0021, and BF_{10}s > 16.64).

Results

In contrast to Brady et al.'s (2013) findings, our results showed that participants exhibited better color memory in the WM than the LTM condition. Using maximum-likelihood estimation, as did Brady et al., we found that all participants, without exception, exhibited more precise memory in the WM condition, $t(22) = 7.77$, $p < .001$, $\eta_p^2 = .73$, 90% CI = [.53, .81], $\text{BF}_{10} = 1.6 \times 10^5$; CBP: $p < .001$, and for 20 out of 23 participants, WM was associated with lower rates of guessing, $t(22) = 4.70$, $p < .001$, $\eta_p^2 = .50$, 90% CI = [.23, .65], $\text{BF}_{10} = 252.15$; CBP: $p < .001$. See Figures 2 and 3 and Table 1 for the maximum-likelihood estimates in both conditions (Table 1 and Fig. 3 also depict our dependent measures estimated using a Bayesian hierarchical model; see details in the Results section of Experiment 2).

Experiment 2

Experiment 2 was devised to test whether the WM fidelity advantage revealed by all Experiment 1 participants would be manifest in a direct replication of the study by Brady et al. (2013, Experiment 1a; see the Procedure and Design section for modifications). Hence, item memory was no longer tested, and the WM condition included three consecutive continuous-color reports, rather than just one. Furthermore, we followed Brady et al.'s analysis scheme, deriving the WM maximum-likelihood estimates from the aggregated reports.

Unlike Brady et al. (2013), we also analyzed each WM report separately to explore the effects of report order on fidelity. If fidelity declines across WM reports (Adam et al., 2017; Peters et al., 2018), then the SD estimate of the first WM report would likely have been better than the aggregate, and the third WM report would likely have been worse. Because Brady et al. found the aggregated SD estimate of WM to be equal to the SD estimate of LTM, then the estimate of LTM would likely have been better than that of the third WM report—an entirely unintuitive result that contradicts the hypothesis of equal WM–LTM fidelity. With regard to the fidelity limit within WM itself, a decline in color precision over reports would suggest that it is not immune to the effects of time-correlated mechanisms and, in particular, to output interference.

Method

Participants. Twenty-four Tel Aviv University students participated in the experiment for course credit (19 women; 22 right-handed; age: M = 22.96 years, SEM = 0.63). Three participants were excluded from the analyses because they met the second predefined exclusion criterion (see the Analysis section of Experiment 1).

Stimuli. The stimulus set comprised the 540 experimental images from Experiment 1 with an additional 26 images for the Brady et al. (2013) repetition-detection

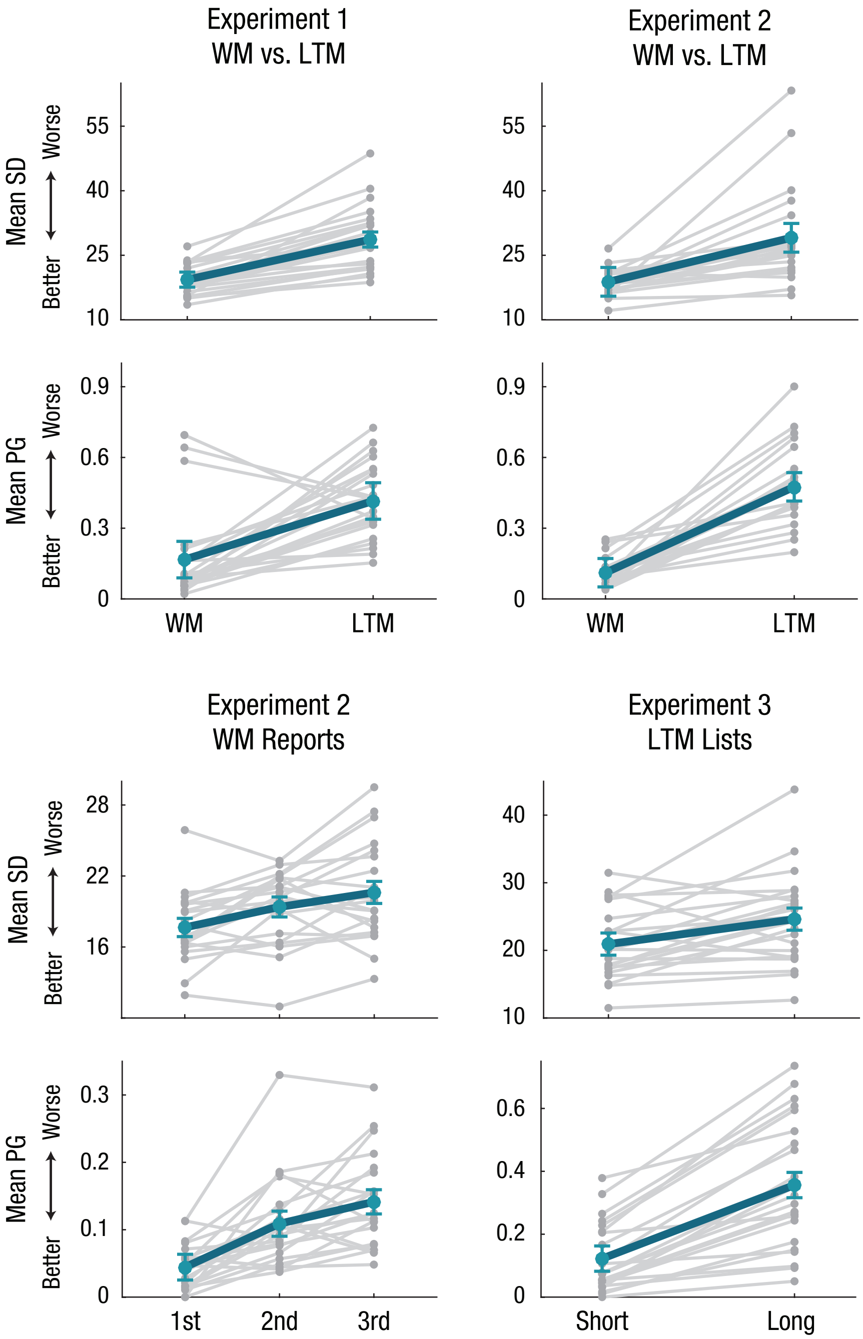

Fig. 2. Color-memory performance (estimated via maximum-likelihood estimation) in Experiments 1 to 3. The top row shows mean fidelity estimates (SD), and the bottom row shows mean probabilities of guessing (PG). For both estimated parameters, better performance is reflected by lower values on the *y*-axis. Gray dots and the thin lines connecting them indicate individual data, and cyan thick dots and lines depict group means for each condition. Error bars denote within-participants 95% confidence intervals (Loftus & Masson, 1994). WM = working memory; LTM = long-term memory.

task in the LTM condition (see the Procedure and Design section below). For each participant, the 540 experimental images were divided into three blocks—two WM blocks and one LTM block—for a total of 180 images per block. The images were counterbalanced such that across participants, each image appeared an equal number of times in each block.

Procedure and design. The first WM and LTM blocks (henceforth, the *replication blocks*) in Experiment 2 were

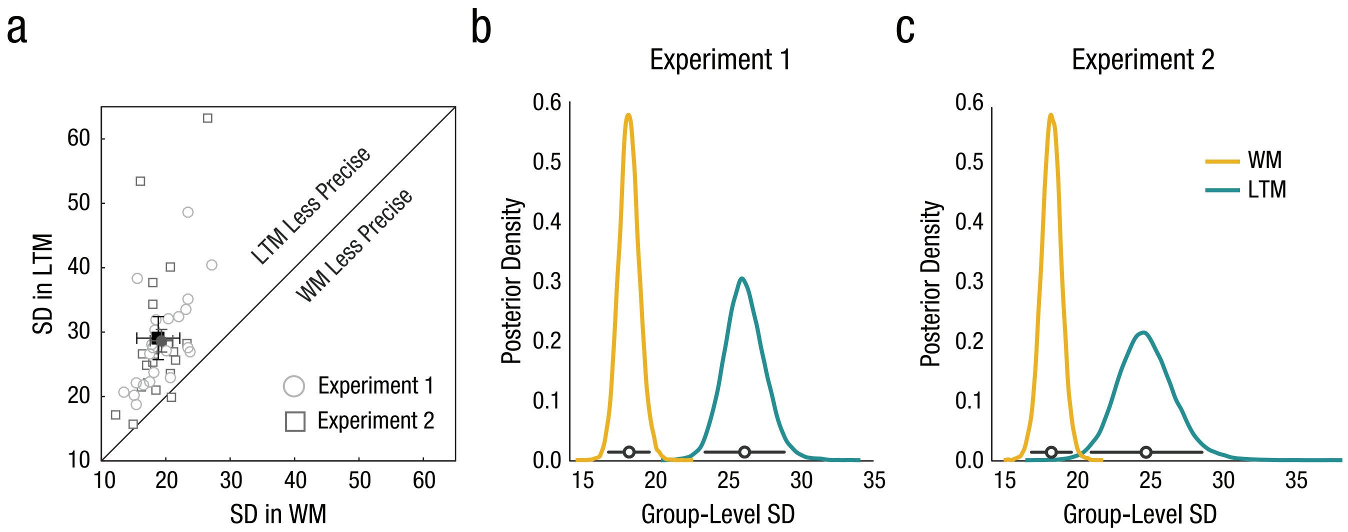

Fig. 3. Fidelity estimates (SDs) for working memory (WM) and long-term memory (LTM) in Experiments 1 and 2. Each shape in (a) represents the SD estimate for an individual participant in Experiment 1 (*N* = 23) and Experiment 2 (*N* = 21), as derived from maximum-likelihood estimation. The diagonal reflects no difference between SD estimates for WM and LTM, shapes above the diagonal reflect better fidelity in WM, and shapes below the diagonal reflect better fidelity in LTM. Darker shapes represent group means, with error bars indicating within-participants 95% confidence intervals (Loftus & Masson, 1994) for SD estimates of both WM (horizontal bars) and LTM (vertical bars). The line graphs depict posterior densities for group-level SD estimates in the LTM and WM conditions in Experiment 1 (b) and Experiment 2 (c). Circles above the *x*-axes indicate means of the posterior distribution, separately for each condition, with error bars indicating 95% highest-density intervals.

a direct replication of these conditions in Experiment 1a of the study by Brady et al. (2013). To improve performance in the WM and LTM conditions, we made the following modifications from the original study. First, to spotlight the WM–LTM comparison, we did not include the perception condition (in which participants matched the color of a visible object). Second, to avoid variations in perceived color space between trials, we used our Experiment 1 unified-hue stimuli. Third, neighboring study items were different by at least 40° (see Section

Table 1. Estimated Fidelity (SD) and Probability of Guessing (PG) in Experiments 1 to 3

Experiment and condition	Maximum-likelihood estimation		Bayesian hierarchical modeling	
	SD	PG	SD	PG
Experiment 1[a]				
LTM	28.66 (1.52)	.42 (.03)	26.18 [23.47, 28.90]	.42 [.35, .49]
WM	19.34 (0.74)	.17 (.04)	18.27 [16.85, 19.65]	.18 [.12, .25]
Experiment 2[b]				
LTM	29.07 (2.51)	.48 (.04)	24.78 [21.01, 28.62]	.48 [.41, .56]
WM	18.84 (0.68)	.11 (.01)	18.28 [16.92, 19.66]	.12 [.09, .14]
First report[c]	17.65 (0.65)	.04 (.01)	17.02 [15.70, 18.27]	.05 [.03, .07]
Second report	19.38 (0.67)	.11 (.01)	18.85 [17.35, 20.34]	.11 [.08, .14]
Third report	20.61 (0.92)	.14 (.02)	19.55 [17.72, 21.36]	.15 [.11, .18]
Experiment 3				
Long list	24.62 (1.40)	.36 (.04)	22.03 [19.36, 24.62]	.36 [.28, .44]
Short list	20.93 (1.08)	.12 (.02)	19.14 [17.04, 21.10]	.13 [.09, .17]

Note: Mean estimates are presented for maximum-likelihood-estimation fitting (with standard errors in parentheses). Group-level mean estimates are presented for the Bayesian hierarchical model (with 95% highest-density intervals in brackets). Higher SD estimates indicate lower fidelity. WM = working memory; LTM = long-term memory.

[a] In Experiment 1, the analysis was conducted only on hit trials. [b] In Experiment 2, the WM–LTM comparison was conducted on the replication blocks. [c] The WM-report comparisons were conducted on data across the two WM blocks.

SM1) to prevent confusion between adjacent items with similar colors.

Block order was counterbalanced across participants. Each block included 180 unique images, for a total of 60 trials in the WM block (from which 180 reports would be obtained) and 180 study and test items in the LTM block.

Because we sought to analyze performance of each WM report separately, we were concerned that 60 WM trials would render the mixture-model parameter estimates less reliable (Lawrence, 2010). Therefore, following the completion of the replication blocks, participants were asked to perform a second, extra WM block. It comprised 180 unique images divided into 60 trials, for a total of 120 observations per WM report across the two WM blocks. For most participants (n = 15), this extra block was administered in a separate session. No significant difference was found between the two groups of participants in any of the color-memory estimates—SD: $t(19) < 1$, BF_{01} = 2.38; PG: $t(19) = 1.19$, p = .249, η_p^2 = .07, 90% CI = [.00, .28], BF_{01} = 1.49. Importantly, data from the extra block were not included in the direct-replication analysis.

The WM and LTM conditions were similar to those in Experiment 1, with the following four exceptions aimed to conform to the original study by Brady et al. (2013). First, studied items alone were tested for color memory. Second, the WM condition included a whole-report procedure in which each of the three studied items was tested sequentially. Test order within a trial was randomized. Third, the assigned color of the study items was not identical for all participants but was rather randomly chosen from a uniform distribution of angles (between 0° and 360° added to the original unified hue) for each participant (see Section SM1). Fourth, the study phase in the LTM condition included Brady et al.'s encoding task. Participants were thus asked to detect a repetition of items, and they were given feedback only when they made a response. The fixation cross turned green for hits (correct press) and red for false alarms (incorrect press). For misses and correct rejections, the fixation cross remained black. The 26 repeated pairs were not included in the analyses. To ensure that participants were alert throughout the entire study phase, we programmed the repeated images to appear in intervals of at least four items.

Analysis. Two main analyses were conducted in the current experiment. First, a comparison between WM and LTM—the replication analysis—was made using a paired-samples t test for each of our dependent measures. Only the first WM block was included in this analysis. For this WM block, responses from the three reports were aggregated into a single response-errors distribution, from which color estimates were obtained. Second, a comparison between the three WM reports was run using a repeated measures ANOVA with WM report as a within-participants variable for each of the two dependent measures. For this analysis, data from the two WM blocks were collapsed for each participant, as no significant difference between these blocks was found in any of the color-memory estimates—SD: $t(20) < 1$, BF_{01} = 3.08; PG: $t(20) = 1.54$, p = .139, η_p^2 = .11, 90% CI = [.00, .32], BF_{01} = 1.58.

Results

A comparison of LTM and WM: the replication analysis. Participants were better in the WM compared with the LTM condition in both color-memory estimates (see Figs. 2 and 3 and Table 1). Specifically, for WM, color memory was more precise, $t(20) = 4.50$, $p < .001$, η_p^2 = .50, 90% CI = [.21, .66], BF_{10} = 139.11, and was less likely to be forgotten, $t(20) = 8.89$, $p < .001$, η_p^2 = .80, 90% CI = [.62, .86], $\text{BF}_{10} = 6.1 \times 10^5$. Indeed, the WM condition induced lower SD estimates for 20 out of 21 participants (see Fig. 3) and lower PG estimates for all participants (CBPs; all ps < .001).

Interestingly, the fidelity advantage of WM over LTM remained significant even when we compared SD estimates in the LTM condition with SD estimates in the third (last) WM report, which was presumably subject to the maximum potential effects of output interference, $t(20) = 3.54$, p = .002, η_p^2 = .39, 90% CI = [.11, .57], BF_{10} = 19.48. Because Brady et al. (2013) did not adopt any exclusion criteria in their analyses, we reanalyzed our data, this time excluding no participants. This did not change any aspect of our results. Most importantly, the SD estimate for LTM (M = 30.77°, SEM = 2.88) was still worse than the aggregated SD estimate for WM (M = 20.12°, SEM = 1.06), $t(23) = 3.97$, $p < .001$, η_p^2 = .41, 90% CI = [.14, .58], BF_{10} = 54.75; CBP: $p < .001$, and for the third WM report (M = 21.23°, SEM = 0.94), $t(23) = 3.61$, p = .001, η_p^2 = .36, 90% CI = [.11, .54], BF_{10} = 25.17; CBP: $p < .001$.

Together, Experiments 1 and 2 failed to replicate the equal WM–LTM fidelity reported by Brady et al. (2013), even in a direct replication of the original design. We next assessed two possible criticisms of our failure to replicate Brady et al.'s hypothesis of equal WM–LTM fidelity.

A trade-off strategy in the LTM condition. The difference between our findings and those of Brady et al. (2013) is centered on LTM performance. Our participants were less precise in LTM than were Brady et al.'s (SD estimates were 30.2° and 29.1° in our Experiments 1 and 2, respectively,[1] compared with 19.3° and 20.3° in Brady et al.'s Experiments 1a and 1b, respectively). Yet our participants also remembered more colors (PG estimates were .49 and .48 in our Experiments 1 and 2, respectively, compared with .58 and .63 in Brady et al.'s Experiments 1a and 1b, respectively). Therefore, one could interpret our failure to replicate the equal WM–LTM fidelity as mediated by a trade-off strategy used by our participants in the LTM condition. Namely, they may have invested greater resources in remembering the colors (low PG estimates) at the cost of color precision (high SD estimates).

To test this interpretation, we divided our participants into two groups on the basis of their PG estimates, generating a low- and a high-PG group for each experiment. In Experiment 1, the mean PG estimate in the low-PG group (n = 12) was .38 (SEM = .03), whereas for the high-PG group (n = 11), the mean was .60 (SEM = .02). In Experiment 2, the mean PG estimate in the low-PG group (n = 10), was .34 (SEM = .02), whereas in the high-PG group (n = 11), it was .60 (SEM = .04). If a trade-off strategy mediated our failure to replicate the equal WM–LTM fidelity, we would expect to observe a significant difference in SD estimates between the two groups. Specifically, the high-PG group should include lower SD estimates (better precision) compared with the low-PG group. This is because high-PG participants would have presumably directed more resources toward color precision at the cost of remembering the color. On examination of the SD estimates, we found that this was the case for neither Experiment 1 nor Experiment 2. SD estimates of both groups remained high (at ~29°) and differed significantly in neither experiment. Specifically, for Experiment 1, the mean SD estimates were 28.90° (SEM = 1.89) and 31.55° (SEM = 3.66) for the low- and high-PG groups, respectively, $t(21) < 1$; BF_{01} = 2.25. For Experiment 2, the mean SD estimates were 28.13° (SEM = 3.31) and 29.92° (SEM = 3.87) for the low- and high-PG groups, respectively, $t(19) < 1$; BF_{01} = 2.45. Importantly, in both experiments, the high-PG groups had similar PG estimates to those observed in Brady et al.'s study, yet their SD estimates were nevertheless considerably higher—signifying worse precision.

Inflation of SD estimates in LTM. Recent studies have shown that SD estimates are artificially inflated when memory performance is poor (Oberauer, Stoneking, Wabersich, & Lin, 2017; Sutterer & Awh, 2016). When guessing rates are high, the von Mises distribution around the study color includes fewer observations, so it is more difficult to obtain accurate SD estimation. Because LTM conditions usually involve higher guessing rates than WM conditions, it is possible that our WM–LTM fidelity effect simply reflects inflated SD estimates in the LTM condition rather than a true difference in fidelity between WM and LTM. Notably, this possibility also applies to the experiments performed by Brady et al. (2013), in which the LTM conditions involved even higher guessing rates (58% and 63% in Experiments 1a and 1b, respectively).

In a recent article, Oberauer et al. (2017) suggested that estimating mixture-model parameters in a Bayesian hierarchical framework (rather than a maximum-likelihood framework) minimizes the SD-estimation bias in conditions of low performance. Indeed, for high guessing rates (PG estimates of .2 or higher), Oberauer and colleagues found the recovery of SD estimates to be very accurate. The advantage of using hierarchical modeling over fitting maximum-likelihood estimates for individual participants is that the former method enables one to compute group-level parameter estimations while still accounting for individual differences. Participants are treated as belonging to a single population (i.e., partial pooling), and individuals' parameter estimates are informed by all other participants' parameter estimates (Gelman et al., 2014). This is especially important when guessing rates are high, and thus the data provide sparse information for SD estimation.

In an auxiliary analysis, we also computed group-level SD and PG estimates using Oberauer et al.'s Bayesian implementation of the Zhang-Luck model. Group-level SD and PG parameters were estimated for each condition separately by computing a posterior distribution of credible parameter values given the data, using the Markov chain Monte Carlo method (see Section SM3 in the Supplemental Material for implementation details and osf.io/93cvs for our modeling scripts). To interpret the results, we computed the mean and the 95% highest-density interval (HDI) of the posterior distribution (each step of the Markov chain) of each group-level parameter. Given our priors and data, the HDIs provide an intuitive probabilistic assessment of one's confidence that the estimated parameter falls within a specific range. To assess the difference between two conditions (e.g., between WM and LTM), we subtracted the group-level parameter of one condition from the other in every step of the Markov chain and then computed the mean and 95% HDI of this difference. If the 95% HDI of the difference excludes zero, it is reasonable to conclude that zero is not considered a credible value and, hence, that the two conditions reliably differ.

In both Experiments 1 and 2, the group-level SD estimates for the LTM condition were indeed lower than the mean of individual estimates obtained using maximum-likelihood model fitting (see Table 1). Nevertheless, group-level SD estimates in the LTM condition remained higher than those in the WM condition (see Table 1 and Fig. 3). We found that in both experiments, the 95% HDIs of the WM–LTM SD difference excluded zero (Experiment 1: M = 7.91°, 95% HDI = [4.87, 10.96]; Experiment 2: M = 6.50°, 95% HDI = [2.51, 10.56]). In Experiment 2, this SD difference remained intact when we compared the LTM condition with the third WM report, which was subject to the maximum potential effects of output interference (M = 5.23°, 95% HDI = [1.09, 9.51]). Overall, our findings suggest that the limit on fidelity in WM is higher than—not equal to—that of LTM.

A comparison of WM reports within a trial. Next, we examined whether the fidelity of WM declines with WM reports. We first estimated our dependent measures using maximum-likelihood estimation. As predicted, both SD and PG estimates increased (signifying worse performance) as a function of WM report order (see Fig. 2 and Table 1). The main effect of WM reports was significant for SD estimate, $F(2, 40) = 12.43$, $p < .001$, $\eta_p^2 = .38$, 90% CI = [.17, .52], $BF_{10} = 337.38$, and for PG estimate, $F(2, 40) = 28.79$, $p < .001$, $\eta_p^2 = .59$, 90% CI = [.40, .68], $BF_{10} = 1.0 \times 10^6$. Of 21 participants, 9 showed a monotonic increase in SD estimate and 13 in PG estimate (CBPs: $p = .004$ and $p < .001$, respectively).

We also confirmed the decline in report performance with a Bayesian hierarchical model. The difference between the third and the first WM reports had 95% HDI, which excluded 0 in both the group-level SD estimate (M = 2.53°, 95% HDI = [0.40, 4.82]) and PG estimate (M = 0.10, 95% HDI = [0.06, 0.14]; see Table 1 for group-level parameter estimates).

The decline of color memory over reports suggests that performance on the whole-report procedure does not accurately reflect that of a partial-report procedure, wherein only a single item from each study array is probed. To test this suggestion, we compared performance of the first WM report (which corresponds to a partial-report procedure) with the aggregated measure of WM (for which data from all reports in both WM blocks were aggregated). For both SD and PG estimates, the first report was significantly better than the aggregated measure—SD estimate: $t(20) = 4.92$, $p < .001$, $\eta_p^2 = .55$, 90% CI = [.26, .69], $BF_{10} = 328.91$; PG estimate: $t(20) = 6.96$, $p < .001$, $\eta_p^2 = .71$, 90% CI = [.46, .80], $BF_{10} = 1.9 \times 10^4$. When this analysis was performed in a Bayesian hierarchical model, the difference between the aggregated measure and the first report was found at the group-level PG estimate (M = 0.05, 95% HDI = [0.02, 0.08]) but showed only a tendency toward a reliable difference for the group-level SD estimate (M = 1.33°, 95% HDI = [−0.57, 3.22]).

Experiment 3

Here, we turned to assess the constraint on fidelity within LTM. Brady et al. (2013) observed equal SD estimates for LTM in lists of 20 to 360 items (Control Experiments 2–4; Brady et al., 2013). However, statistical power for these results may have not been sufficient. In Experiment 3, we compared color memory across short lists (18 items) and long lists (180 items). If all mnemonic information is subject to the effects of time-correlated mechanisms, then the short lists—less affected by interference from studied and retrieved items and, perhaps, decay—should result in better color memory than long lists. Experiment 3 was preregistered (osf.io/ksr4d).

Method

Participants. Twenty-four Tel Aviv University students participated in the experiment for course credit or payment (~$10 per hour; 14 women; 24 right-handed; age: M = 25.46 years, SEM = 0.59). One participant was excluded from the analyses because she met the second predefined exclusion criterion (see the Analysis section of Experiment 1 and osf.io/ksr4d).

To maximize the effect size of the current experiment, we followed Benjamini and Hochberg's (1995) procedure. Our sampling plan allowed only one iteration of observing the data. Accordingly, if the results were not significant after the first 24 participants, we planned to run an additional 18 participants, while correcting for false-discovery rate for our particular sampling plan (see the preregistered Method section at osf.io/ksr4d). Our results did not necessitate running an additional 18 participants, yet the p value for which we compared our results was corrected to .04.

Stimuli. The stimuli set consisted of 360 images from the 540 images used in Experiments 1 and 2. The 360 images were assigned to the two experimental conditions (180 items each) and counterbalanced such that across participants, each image appeared an equal number of times in the two experimental conditions. Within each condition, the order of images was random and divided into lists according to the experimental condition (e.g., ten 18-word lists for the short-list condition). In both conditions, the color of the studied items was randomly chosen from a uniform distribution of angles (see Section SM1).

Design and procedure. Our design included a single within-participants variable: list length (short vs. long), with the two dependent measures for color memory. The short-list condition included ten 18-item lists, and the long-list condition included a single 180-item list. Each of the two conditions was blocked, such that all lists of the same condition were presented one after the other. The order of the two experimental conditions was counterbalanced across participants. Because the experiment included only LTM conditions, we added monetary incentive to improve performance. Participants were told that one individual with the best results would be awarded a monetary prize comparable to \$100. The award was granted to the individual with the best fidelity estimate across all experimental conditions.

Before beginning the experimental session, participants underwent a short practice block. At the beginning of each condition, participants were notified of the length of the upcoming lists. Each list contained a study phase followed by an immediate test phase (see Fig. 1), both identical to the LTM condition of Experiment 2 with the exception that the LTM encoding task in Experiment 3 was the real-life color-judgment task used in Experiment 1. At the end of each study-test sequence, participants were asked to continue to the next list when ready by pressing the space bar.

Results

Color memory was better in the short-list condition compared with the long-list condition for both SD and PG estimates (see Fig. 2 and Table 1). The short list induced better precision of color for 18 of 23 participants, $t(22) = 3.29$, $p = .003$, $\eta_p^2 = .33$, 90% CI = [.08, .52], $BF_{10} = 12.45$; CBP: $p = .005$, and lower probability of guessing the color, $t(22) = 8.50$, $p < .001$, $\eta_p^2 = .77$, 90% CI = [.58, .84], $BF_{10} = 6.4 \times 10^5$, for all participants (CBPs: $p < .001$). Thus, we applied the same maximum-likelihood-estimation analysis scheme as did Brady et al. (2013), and our findings failed to replicate their results (Control Experiments 2–4), wherein the same SD estimate was observed for short and long LTM lists.

When estimated in a Bayesian hierarchical model, the difference between the two lists was evident in the group-level PG estimate ($M = .23$, 95% HDI = [.14, .32]) but only showed a tendency toward a reliable difference for the group-level SD estimate ($M = 2.89°$, 95% HDI = [−.53, 6.13]; see Table 1).

Although only a mild fidelity effect was observed in the LTM analysis when applying the Bayesian hierarchical model, we remind readers that our motivation in running this model was only to control for a concern in the maximum-likelihood-estimation procedure. That concern centered on a potential upward bias in SD estimates when guessing rates are high, which is usually the case in LTM conditions. Importantly, keeping to the maximum-likelihood analysis, we applied an additional control procedure to deal with the concern of an upward bias in SD estimation (Oberauer et al., 2017; Sutterer & Awh, 2016). Specifically, we simulated data by varying the true guessing rates and used maximum-likelihood-estimation fitting to locate the maximum guessing probability (i.e., PG cutoff), after which a systematic bias in SD estimation occurred. Participants with PG estimates above this cutoff were then excluded from the analyses (for details, see Section SM4 in the Supplemental Material and our simulation script at osf.io/93cvs). Using this procedure, we replicated the upward SD-estimation bias from a PG estimate of .65. Importantly, we found a significant fidelity effect for list length, even when excluding the additional 2 participants who met the PG cutoff (SD estimates in the short-list condition: $M = 20.34°$, $SEM = 1.06$; and in the long-list condition: $M = 24.76°$, $SEM = 1.51$), $t(20) = 4.03$, $p < .001$, $\eta_p^2 = .45$, 90% CI = [.16, .62], $BF_{10} = 52.29$. An even stronger pattern was observed when the PG cutoff was set to .6, as observed by Sutterer and Awh (2016), and resulted in the removal of 4 participants, $t(18) = 4.98$, $p < .001$, $\eta_p^2 = .58$, 90% CI = [.28, .71], $BF_{10} = 285.59$. Parenthetically, using this alternate procedure did not change any of the effects reported in this article (all ps < .0015, and all BF_{10}s > 29.55). On balance, therefore, we interpret our results to suggest better fidelity for the short list compared with the long list.

General Discussion

Brady et al. (2013) proposed that the fidelity plateau is a general property of memory. This notion arose from the finding of equal limits on fidelity in WM and LTM. Here, we showed that color fidelity was stable neither across nor within memory systems. We used both maximum-likelihood estimation and Bayesian hierarchical modeling and found that the estimated fidelity was better in WM than in LTM. In WM, fidelity worsened across reports in the whole-report procedure. In LTM, long lists had poorer fidelity than short ones. Our findings suggest that fidelity is not immune to the effects of time-correlated mechanisms (e.g., output interference and, perhaps, decay) and is no different from other types of mnemonic information. Thus, the justification for a general fidelity constraint in memory seems to no longer be valid.

If they do not share the same fidelity limit, what is the relation between WM and LTM precisions? We assume that information moves downstream from WM to LTM. This idea can be conceptualized in terms of WM as a separate memory store (Atkinson & Shiffrin, 1968) and as a distinct state of LTM representations (Cowan, 2005; Oberauer, 2002). In either case, representations in WM enjoy more accessibility and can be easily read out with higher accuracy because, compared with LTM, time-correlated mechanisms have less opportunity to affect them. Consequently, when moving from WM to LTM, information either can remain intact—yielding equal fidelity—or may be weakened—yielding worse fidelity. What is not possible is to find an improvement in fidelity when moving from WM to LTM. Accordingly, we suggest that the worst fidelity in WM dictates the best potential fidelity in LTM. This notion is supported by the similar fidelity estimates observed in our LTM condition, which was least subject to time-correlated effects (short list, Experiment 3), and our WM condition, which was most subject to time-correlated effects (third report, Experiment 2), $t(42) < 1$, $\text{BF}_{01} = 3.29$; group-level SD-estimate difference: $M = -0.42°$, 95% HDI = [−3.12, 2.32]. Thus, WM and LTM can potentially yield identical SD estimates. Yet it does not follow that they are limited by the same bound.

Notably, when concluding that the fidelity of WM is better than LTM, fidelity was estimated using the Zhang-Luck mixture model, as was used by Brady et al. (2013). We acknowledge that had we used other mixture models, we could have observed different results. For example, several studies have pointed out the role of categorical representations in continuous-report paradigms in WM (Bae, Olkkonen, Allred, & Flombaum, 2015; Donkin, Nosofsky, Gold, & Shiffrin, 2015; Hardman, Vergauwe, & Ricker, 2017) and in LTM (Persaud & Hemmer, 2016). Accordingly, it is possible that when the active maintenance of mnemonic representations is interrupted, participants rely more on categorical representations than on continuous ones (Hardman et al., 2017). Because categorical responses are not taken into consideration when response errors are computed (from which SD and PG parameters are estimated), our findings may reflect more reliance on categorical representations as a function of time-correlated mechanisms, rather than decline in fidelity per se. Note that we did not ask participants to provide verbal labels of color categories (e.g., Donkin et al., 2015), fearing that this would amplify categorization encoding strategies. Future studies, using controlled color experiments and mathematical modeling, can test whether categorical representations underlie the fidelity disadvantage in LTM. Critically, however, even if we were to find a categorical component in our data, it would still need to be established that this component did not likewise mediate performance in the work by Brady et al.

To conclude, our study provides substantial evidence that the precision of memory follows the dynamics of memory for the event itself. It declines as a function of time-correlated mechanisms both across and within memory systems.

Action Editor

Caren Rotello served as action editor for this article.

Author Contributions

N. Biderman, R. Luria, and Y. Goshen-Gottstein designed the experiments. A. R. Teodorescu assisted with programming. N. Biderman and R. Hajaj collected the data. N. Biderman analyzed the results. N. Biderman and Y. Goshen-Gottstein wrote the manuscript. All the authors approved the final manuscript for submission.

Acknowledgments

We thank Dan Biderman for his insightful comments on the manuscript and for his help in implementing the Bayesian hierarchical model.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This study was supported by Israel Science Foundation Grant No. 1922/15 (to Y. Goshen-Gottstein).

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618813538

Open Practices

Data from all three experiments, a codebook, data-analysis scripts, and hue-unification, Bayesian-modeling, and simulation scripts have been made publicly available on the Open Science Framework (OSF) and can be accessed at osf.io/93cvs. Experiment 1 was exploratory and thus was not preregistered. Experiment 2 was not formally preregistered, yet because it was a direct replication of Experiment 1a by Brady, Konkle, Gill, Oliva, and Alvarez (2013), it followed the same procedure. Experiment 3 was preregistered on the OSF and can be accessed at osf.io/ksr4d. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618813538. This article has received the badges for Open Data and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Note

[1] To allow a direct comparison of our Experiment 1 with Experiment 1a of Brady et al. (2013), in which item memory was not tested, we combined hits and misses and then computed SD and PG estimates.

References

Adam, K. C., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, *97*, 79–97.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*, 415–445.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W.

Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 90–195). New York, NY: Academic Press.

Bae, G. Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*, 744–763.

Bays, P. M., Catalao, R. F., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), Article 7. doi:10.1167/9.10.7

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*, 851–854.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B: Methodological*, *57*, 289–300.

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), Article 4. doi:10.1167/11.5.4

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, USA*, *105*, 14325–14329.

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science*, *24*, 981–990.

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

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*, 539–576.

Cowan, N. (2005). *Working memory capacity*. New York, NY: Psychology Press.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review*, *112*, 3–42.

Donkin, C., Nosofsky, R., Gold, J., & Shiffrin, R. (2015). Verbal labeling, gradual decay, and sudden death in visual short-term memory. *Psychonomic Bulletin & Review*, *22*, 170–178.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton, FL: CRC Press.

Hardman, K. O., Vergauwe, E., & Ricker, T. J. (2017). Categorical working memory representations are used in delayed estimation of continuous colors. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 30–54.

Hollingworth, A. (2004). Constructing visual representations of natural scenes: The roles of short- and long-term visual memory. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 519–537.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299.

Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*(8). doi:10.18637/jss.v020.i08

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, Article 863. doi:10.3389/fpsyg.2013.00863

Lawrence, M. A. (2010). Estimating the probability and fidelity of memory. *Behavior Research Methods*, *42*, 957–968.

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge, England: Cambridge University Press.

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476–490.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for Common Designs (Version 0.9.11-1) [Computer software]. Retrieved from https://github.com/richarddmorey/BayesFactor/tree/0.9.11-1

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention.*Journal of Experimental Psychology: Learning, Memory, and Cognition*,*28*, 411–421.

Oberauer, K., & Lin, H. Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*, 21–59.

Oberauer, K., Stoneking, C., Wabersich, D., & Lin, H. Y. (2017). Hierarchical Bayesian measurement models for continuous reproduction of visual features from working memory. *Journal of Vision*, *17*(5), Article 11. doi:10.1167/17.5.11

Persaud, K., & Hemmer, P. (2016). The dynamics of fidelity over the time course of long-term memory. *Cognitive Psychology*, *88*, 1–21.

Peters, B., Rahm, B., Czoschke, S., Barnes, C., Kaiser, J., & Bledowski, C. (2018). Sequential whole report accesses different states in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 588–603.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Richter, F. R., Cooper, R. A., Bays, P. M., & Simons, J. S. (2016). Distinct neural mechanisms underlie the success, precision, and vividness of episodic memory. *eLife*, *5*, Article e18260. doi:10.7554/eLife.18260

Sadeh, T., Ozubko, J. D., Winocur, G., & Moscovitch, M. (2016). Forgetting patterns differentiate between two forms of memory representation. *Psychological Science*, *27*, 810–820.

Schurgin, M. W., & Flombaum, J. I. (2015). Visual long-term memory has weaker fidelity than working memory. *Visual Cognition*, *23*, 859–862.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569.

Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, *13*(10), Article 9. doi:10.1167/13.10.9

Sutterer, D. W., & Awh, E. (2016). Retrieval practice enhances the accessibility but not the quality of memory. *Psychonomic Bulletin & Review*, *23*, 831–841.

Talmi, D., Grady, C. L., Goshen-Gottstein, Y., & Moscovitch, M. (2005). Neuroimaging the serial position curve: A test of single-store versus dual-store models. *Psychological Science*, *16*, 716–723.

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*, 49–60.

Van den Berg, R., & Ma, W. J. (2014). "Plateau"-related summary statistics are uninformative for comparing working memory models. *Attention, Perception, & Psychophysics*, *76*, 2117–2135.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), Article 11. doi:10.1167/4.12.11

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*, 233–235.

Zhang, W., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological Science*, *20*, 423–428.