A Replication of “Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items” by Timothy F. Brady & George A. Alvarez (2011, Psych Science)

Michael L. Waskom
Michael C. Frank
Stanford University

Introduction

Brady and Alvarez presented a study of visual short-term memory (VSTM) in which the central finding is that item memory can be biased by the ensemble statistics of a visual display. Specifically, using a grouping principle of color, memory for the size of a single circle stimulus was biased towards the mean of other stimuli in that group. In other words, for a specific display in which red circles were relatively larger than blue circles, the same-sized circle would be remembered as larger if originally presented in red than if presented in blue. Using a within-subjects design, Brady and Alvarez calculated a measure of bias towards the group mean for size reports. Additionally, they presented a simple probabilistic model encoding stimulus information at multiple levels of abstraction, which provided a strong fit to the observed empirical data and formalized the computational account of memory processes leading to the main behavioral finding.

Methods

Power Analysis

Retrospective power analyses indicated that the original experiment had an observed power of 0.97. We then performed a prospective power analysis to determine required sample size for the replication experiment. Required *n*s to achieve .8, .9, and .95 power were, respectively, 11, 16, and 19.

Planned Sample

Closely replicating the original procedures, we collected data from 20 workers on Amazon Mechanical Turk. Workers received $0.25 for completing the study.

Materials and Procedure

The replication project followed an identical experimental approach to the original paper (see quoted methods below). As in the original experiment, the same 30 displays were used for each participant. However, we pseudorandomly generated a new set of displays using the method described below, so the stimuli did not exactly match what participants in the original experiment encountered. Additionally, the exact colors used in the original experiment are not reported; we used the HTML colors “OrangeRed”, “LimeGreen”, “RoyalBlue” in an attempt to roughly balance luminance across the colors.

“All observers were presented with the same 30 displays consisting of three red, three blue, and three green circles of varying size and were told to remember the size of all of the red and blue circles, but to ignore the green circles. We included the green distractor items in the displays because we believed they would encourage observers to encode the items by color, rather than to select all of the items into memory at once. The order of the 30 displays was randomized across observers. Each display appeared for 1.5 s and was followed by a 1-s blank, after which a single randomly sized circle reappeared in black at the location that a red or blue circle had occupied. Observers had to slide the computer mouse up or down to resize this new black circle to the size of the red or blue circle they had previously seen at that location; they then clicked to lock in their answer and start the next trial.

“The nine circles appeared on a gray background that measured 600 × 400 pixels. Each circle was positioned at a random location within an invisible 6 × 4 grid; jitter of ±10 pixels was added to the circles’ locations to prevent collinearities. The size and resolution of observers’ computer monitors were not controlled. However, all observers attested to the fact that the displays were visible in their entirety. Moreover, the critical comparisons are within subjects, and individual differ- ences in absolute size of the displays are factored out by focusing on within-subjects comparisons between conditions.

“Circle sizes were drawn from a separate normal distribution for each color. The mean diameter for the circles of a given color was chosen uniformly on each trial from the interval (15 pixels, 95 pixels), and the diameter of each individual circle was then chosen from a normal distribution with this mean and a standard deviation equal to one eighth of this mean....

“So that we could directly test the hypothesized bias in reported size, we generated 15 matched pairs of displays. First, 15 displays were generated as described; then, another 15 were created by switching the color of the to-be-tested item to the other nondistractor color (and making the reverse switch for another circle, so that there would still be three red circles and three blue circles in the display). Thus, the displays in each of the pairs were matched in the size of all of the circles present and differed only in the color of two circles, including the circle that would later be tested. The 30 displays were randomly interleaved, with the constraint that paired displays could not appear one after the other....”

An example display is shown below. This display was also used as an example on the instruction page of the AMT HIT. Note that it did not appear in the experimental stimuli, although it was generated with the same procedure.

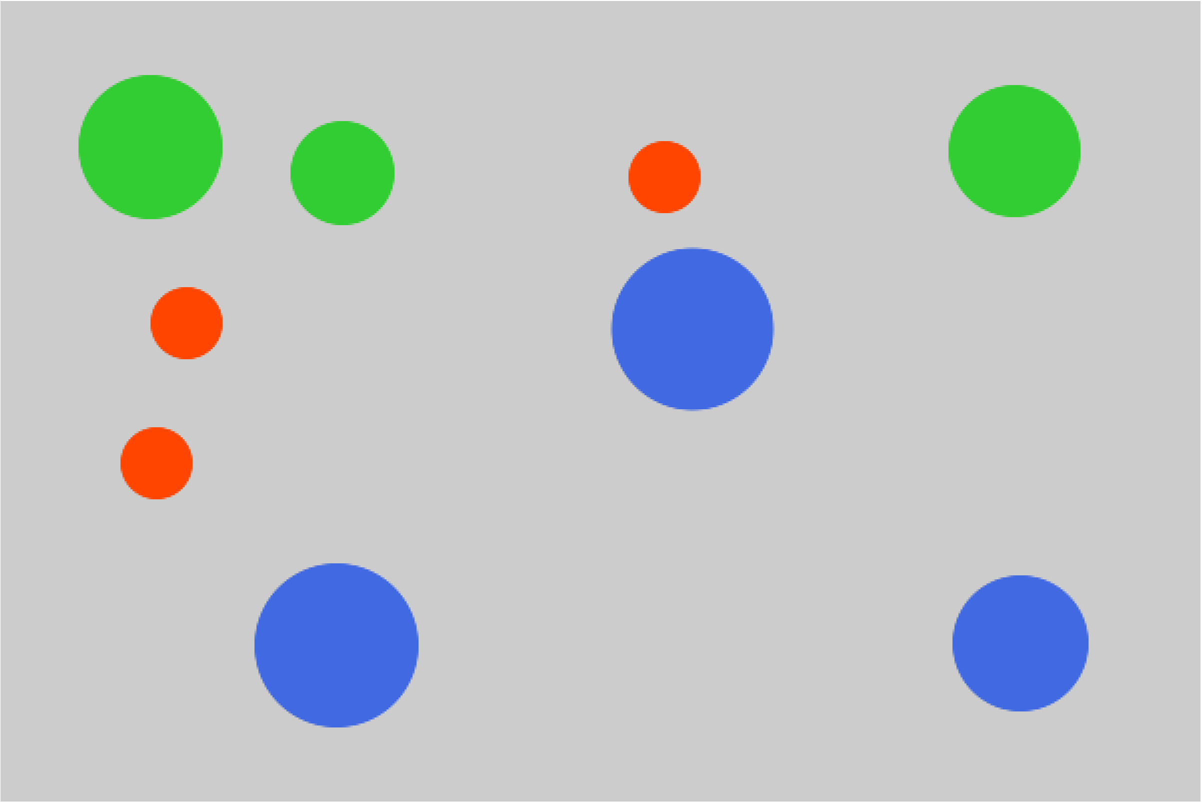

Figure 1: Example stimulus display.

Analysis Plan

The main behavioral analysis we intended to replicate involved comparing the reported size of the target circle as a function of whether the mean size of the same-colored circles was relatively larger than that for the other non-distractor circles on each trial. The measure of this bias was calculated by dividing the reported size for a stimulus grouped with relatively larger circles by the reported size for the complementary condition (i.e., an effect size of 1 indicates no bias, and an effect size > 1 indicates a bias in line with the hypotheses about ensemble statistics). To quantify the significance of the bias, the authors performed a one-sample T-test against the null hypothesis of no bias; based on the degrees of freedom reported for this test it appears to have been performed across subjects after collapsing over displays. The original article also reports the number of subjects with a ratio numerically above 1. We plan to directly replicate this set of analyses.

Because participants could show memory errors such that they accurately report the size of nearby, but differently sized, circle to the one cued on any particular trial, it will be important to control for such outliers, which could lead to inflated or masked estimates of working memory bias. We will exclude individual trials on which the reported size is outside of 3 SDs from the group mean for that display. Additionally, it is possible that some participants will not understand the task, or will not make a full effort. We will exclude at the subject level workers whose overall performance (measured as the mean absolute difference in size between target and report) is outside of 3 SDs from the group mean performance. In practice, we will exclude at the subject level before excluding at the trial level. These exclusion criteria are independent of the main experimental analysis and thus unlikely to bias our results.

Differences from Original Study

As discussed above, although we followed procedures for generating our experimental stimuli identical to those procedures used in the original experiment, we generated them independently and thus the exact sizes, arrangement, and targets are different than what were originally used due to randomness. Although the original paper does not formally present an item analysis, Figure 3 of the paper plots the bias for each display collapsed across participants. This figure reveals a moderate variance in the bias across displays, which could potentially lead to differences in the observed bias with a different sample. Note that the variance appears unrelated to the size of the target circle, but may be related to the difference in group means, which is not reported in the original paper.

Additionally, the procedure for generating stimuli as described in the original paper allows for the possibility of overlapping circles in the displays. We added a constraint such that displays were required to have at least 5 pixels of separation between all circles; it is unknown whether the original stimulus set included a similar constraint or if, given the particular random sample of stimuli originally used, such a constraint was unnecessary. We do not expect this (possible) change to have any influence on the results.

(Post Data Collection) Methods Addendum

Actual Sample

We collected data from 20 unique workers on Amazon Mechanical Turk. Data from one worker were excluded based on the by-participant exclusion rules explicated above; this worker's mean absolute error was 3.47 SDs greater the overall mean error. Additionally, we excluded 6 pairs of trials overall from the analysis based on the by-trial exclusion rule; no more than one pair of trials was excluded for any the participants included in the analyses.

Differences from pre-data collection methods plan

None.

Results

Data preparation

For the main analyses we followed the data treatment outlined by Brady and Alvarez and described above. After excluding outlier trials, we examined overall performance and bias. We measured performance within participants by comparing the actual mean absolute difference between actual and reported sizes against a chance score computed by permuting the reported sizes 1000 times and taking the grand mean of the resulting error distribution. We next computed a bias measure for each pair of displays by dividing the reported size when the target circle was grouped with relatively larger circles by the reported size for the complementary trial. The main conclusions about VSTM were drawn by averaging the bias across displays within each subject and then testing the group mean against 1 (i.e. the value indicating a lack of systematic bias). We additionally explored the effects at the item level and as a function of the difference in group mean sizes, as described and reported in the Exploratory Analyses section.

Confirmatory analysis

The average error collapsed across displays and participants was 11.5 pixels (SD across participants = 4.1 pixels). The chance error score was 25.9 pixels, which was significantly larger than the observed performance as determined with a paired T test (t(18) = -14.83; p = 7.8e-12).

Across subjects, the average bias measure was 1.17 (S.D. across participants = 0.14), indicating a bias towards reporting circles grouped with larger circles as relatively larger. This measure was significantly greater than 1 (t(18) = 5.08; p = 3.9e-5; bootstrapped 95% CI: [1.11, 1.23]) with 17 out of 19 participants numerically above 1. The maximum possible bias (determined as the average ratio of group means across displays) was 1.75, indicating that participants reported sizes that were on average 22% between the actual values and the maximum possible bias.

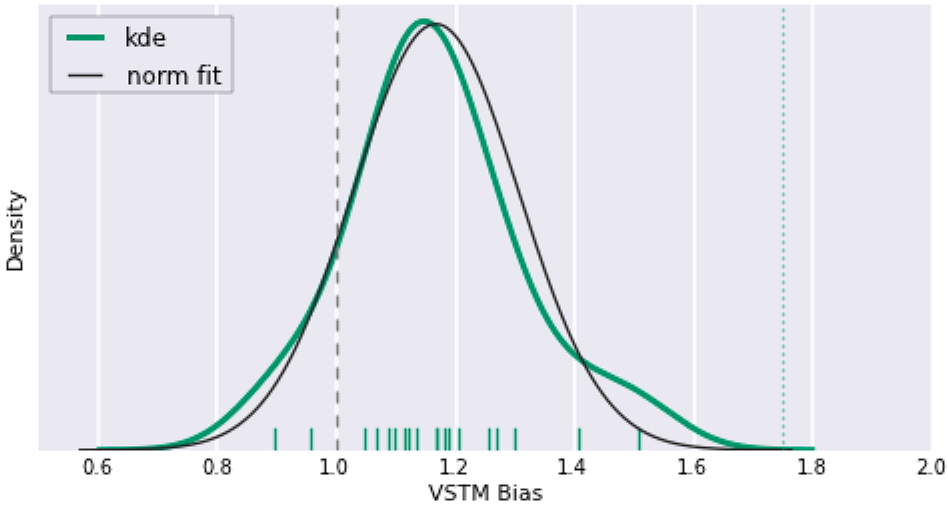

Figure 2: Distribution of VSTM bias across participants. Each vertical tick represents one participant’s mean bias measure. The solid green curve shows a kernel density estimate of the distribution, while the solid black curve shows a maximum likelihood normal fit to the data. The gray dashed line is placed at 1, which indicates no bias, while the dotted line is placed at the location of the maximum possible bias (obtained if participants reported the group means on each display).

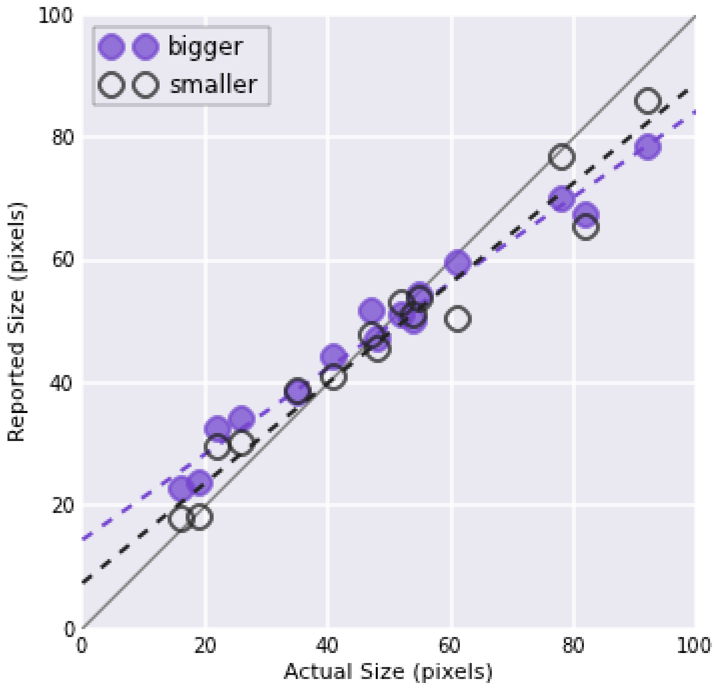

Figure 3: Mean reported size for each target circle plotted against the actual size as a function of whether the target circle was in the relatively larger or smaller color group.

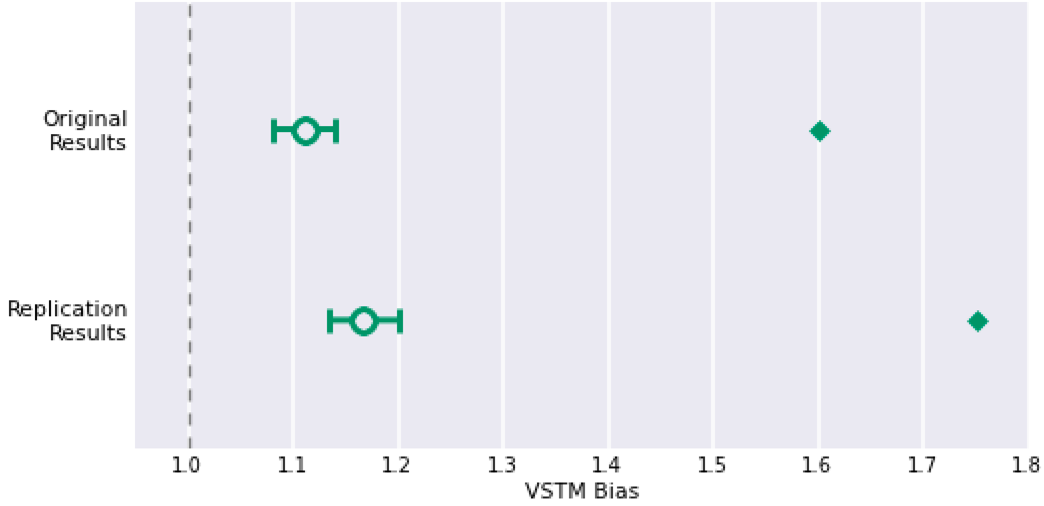

Figure 4: Comparison of original and replication results. Error bars show +/- 1 S.E. across subjects. Diamonds represent the maximum possible bias.

Exploratory analyses

We subsequently tested whether the finding of VSTM bias was robust to the variance across displays. After collapsing the bias measure across subjects, we found that the mean bias was significantly different from 1 [SD = 0.22, *t*(14) = 3.03, *p* = 0.0045, 95% CI: (1.07, 1.28)] with average responses on 14 out of the 15 pairs of displays showing the effect. As the distribution of display biases appeared non-normal (skewed right), we also measured the sample median and its 95% CI, which did not cover 1 [median = 1.08; 95% CI (1.03, 1.32)].

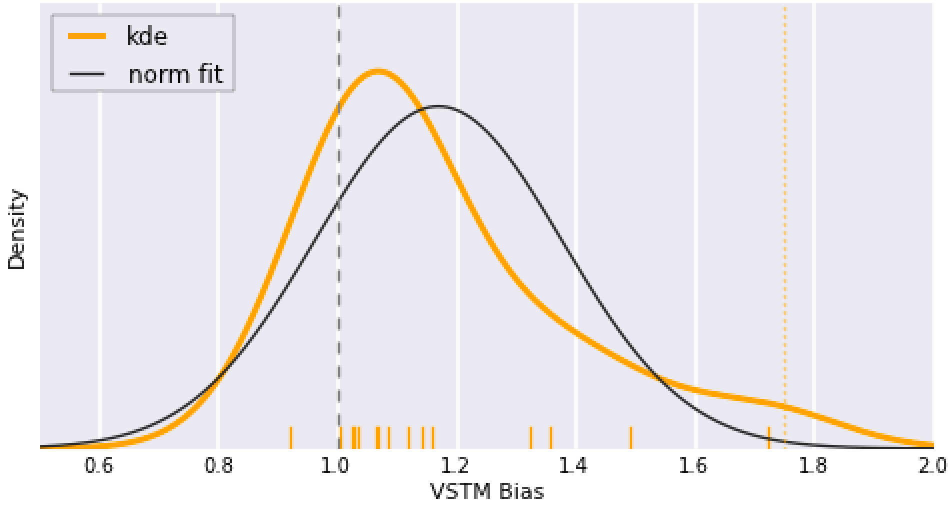

Figure 5: Distribution of VSTM bias across displays. Conventions are as in Figure 2.

We additionally tested whether VSTM bias was influenced by the difference in mean sizes of the each group of circles. For this analysis, we fit a linear mixed model predicting bias with the ratio of group means and fitting both random intercepts and slopes by subject. In this model, the regression coefficient for the group mean ratio was significantly different from 0 as determined by a likelihood ratio test of nested models (β = 0.25; S.E. = 0.055; *t* = 4.62; *p* = 0.0027).

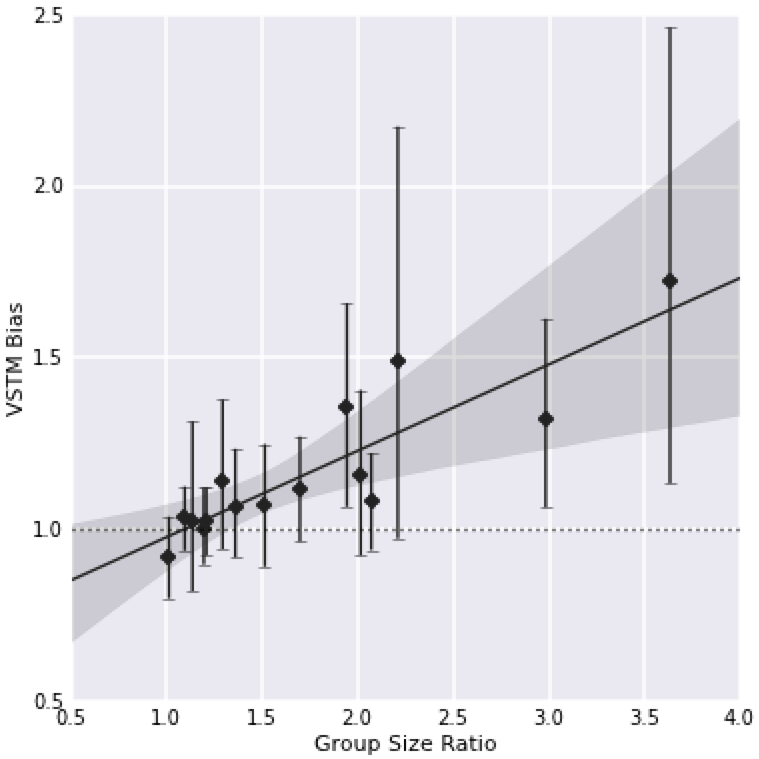

Figure 6: VSTM bias as a function of the ratio in group means (mean of relatively larger color divided by mean of relatively smaller color). Each point represents the mean bias for one display with error bars showing the bootstrapped 95% CI for the mean, while the shaded error band on the regression line show its bootstrapped 95% CI.

Discussion

Summary of Replication Attempt

Our experiment resulted in a full replication of the original results by Brady and Alvarez. To restate the main differences between our experiment and that reported in the original paper, our results include both a different random sample of AMT workers as well as a different random sample of possible displays. Our empirical finding was that VSTM reports are biased by ensemble statistics such that, when a grouping principle is salient, memory for an individual item will be pulled towards the central tendency of the group it belongs to. From a theoretical perspective, these results stand in contradiction to “slot” models of VSTM that conceive of memory representations for items as independently maintained and uninfluenced by other contents of memory.

Commentary

We additionally performed an item analysis to test whether the distribution of bias measures with respect to displays was centered on 1. Although this distribution was right-skewed, it did not appear to be centered on 1. A subsequent analysis showed that the average bias measured for each display was linearly related to the ratio between the means of the two groups. In theory, the difference between two groups is distributed as an exponential random variable with location parameter at 1. Other factors, including the precise arrangement of the stimuli and the difference between the size of the target circle and the mean of its group, likely influence memory bias but were not analyzed here. In summary, these results provide additional support for the original finding and offer evidence that it was driven neither by idiosyncratic features of the participant sample nor by a set of stimuli that unusually encouraged biased VSTM reports.