On the Category Adjustment Model: Another look at Huttenlocher, Hedges, and Vevea (2000)*

Sean Duffy

Rutgers University-Camden, Department of Psychology

John Smith

Rutgers University-Camden, Department of Economics

Address correspondence to:

John Smith
Department of Economics
Rutgers University-Camden
311 N. 5th Street
Camden, NJ
08102 USA
smithj@camden.rutgers.edu

July 9, 2019

* We thank Roberto Barbera, Alex Brown, I-Ming Chiu, Caleb Cox, L. Elizabeth Crawford, Steven Gussman, Johanna Hertel, Matt Jones, Richard McLean, Rosemarie Nagel, Adam Sanjurjo, and Barry Sopher for helpful comments. This project was supported by Rutgers University Research Council Grants #202297 and #18-AA-00143. John Smith thanks Biblioteca de Catalunya.

Abstract

Huttenlocher, Hedges, and Vevea (2000) (Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, 129, 220-241) introduce the category adjustment model (CAM). Given that participants imperfectly remember stimuli (which we describe as "targets"), CAM holds that participants maximize accuracy by using information about the distribution of the targets to improve their judgments. CAM predicts that judgments will be a weighted average of the imperfect memory of the target and the mean of the distribution of targets. Huttenlocher, Hedges, and Vevea (2000) report on three experiments and conclude that CAM is "verified." We attempt to replicate the conditions in Experiment 3 from Huttenlocher et al. (2000). We analyze judgment-level data rather than averaged data. We find evidence of a bias toward a set of recent targets rather than a bias toward the running mean. We do not find evidence learning. The judgments in our dataset are not consistent with CAM. We discuss other defects in HHV – including dividing by zero. It seems that evidence for CAM is a statistical illusion that appears when researchers analyze data averaged across trials and do not consider a recency bias.

Keywords: judgment, memory, category adjustment model, central tendency bias, recency effects, Bayesian judgments

1. Introduction

Psychologists have understood that judgments are a topic worthy of study, notably because perception and memory are imperfect. One clever method for studying judgments is to present participants with stimuli that have objectively measurable properties. For example, the stimulus could be a line with certain dimensions. We refer to a specific stimulus as a *target*. The target then disappears and participants are asked to reproduce some aspect (length of a line, shade of a color, etc.) of the target. We refer to this as the *response*. This task is repeated for targets of various characteristics.

It has been known for some time that when participants perform repeated judgments tasks there is a bias toward the mean of the distribution of targets (Hollingworth, 1910; Poulton, 1979). For instance, in the judgment of the length of lines, longer lines tend to be underestimated and shorter lines tend to be overestimated. This effect is sometimes referred to as the *central tendency bias*.[1]

Because participants imperfectly remember and perceive the targets, Huttenlocher, Hedges, and Vevea (2000), hereafter referred to as HHV, propose that participants use information about the distribution of the targets to improve their judgments. HHV name this the *category adjustment model*, hereafter referred to as CAM. CAM predicts that judgments will be a weighted average of the imperfect memory of the target and the mean of the distribution of targets.[2] In the description of CAM, the authors state,[3] "Our model is a precisely specified

[1] This is also sometimes referred to as the *regression effect* (Stevens and Greenbaum, 1966). The representativeness heuristic (Kahneman and Frederick, 2002; Kahneman and Tversky, 1973) makes similar predictions. Crosetto et al. (2019) find evidence of the central tendency bias in responses to belief elicitations when the distribution is known to be uniformly distributed.

[2] HHV refer to this as the *category*.

[3] The entire paragraph is as follows: Our model is a precisely specified Bayesian model. It holds that in pursuing the goal of maximizing accuracy, people use prior information in estimating stimulus values that are represented

Bayesian model…The mean of this posterior distribution is called the Bayesian estimate; it has the property of being the 'most accurate' estimate in the sense that it minimizes average error." (p. 221). Since judgments will be an optimal weighted average of the imperfect memory of the target and the mean of the distribution, CAM offers a Bayesian explanation of the central tendency bias.

In order to test the predictions of CAM, HHV perform three experiments. Participants perform a series of judgment tasks on the fatness of computer generated images of fish (Experiment 1), the greyness of squares (Experiment 2), and the lengths of lines (Experiment 3). In each of these experiments, participants complete these judgments under four different distributions of targets, which exhibit different means and standard deviations. HHV conduct their analyses on data that had been averaged across trials and averaged across sets of previous targets. HHV conclude by stating, "The experiments verified that people's stimulus estimates are affected by variations in a prior distribution in such a manner as to increase the accuracy of their stimulus reproductions"[4] (p. 220).

However, despite the assertion of HHV to the contrary, one simple alternate hypothesis is that there is a bias toward a set of recent targets rather than a bias toward the mean of the distribution. We note that this is a non-Bayesian explanation because participants are not predicted to exhibit learning.

---

inexactly. Prior information is incorporated into decision making in the form of an explicit prior distribution, and the inexactness of the fine-grained information is incorporated as a sampling distribution. Given a category (an explicit prior distribution) and an inexact stimulus value (a sampling distribution describing the uncertainty of current data), Bayes's theorem provides a method for combining the information to provide estimates with certain optimal properties. A posterior distribution summarizes uncertainty after combining the uncertain data and the prior information. The mean of this posterior distribution is called the Bayesian estimate; it has the property of being the "most accurate" estimate in the sense that it minimizes average error.

[4] We also note that "verified" is a word that appears to be inconsistent with Bayesian inference following an experiment with a limited number of participants performing judgments on a limited set of stimuli.

CAM asserts that participants have beliefs of the distribution of targets. While we are not able to observe the mean of the beliefs the distribution, well-known results show that, under mild assumptions, Bayesian learners will have beliefs that converge to the truth (Savage, 1954; Blackwell & Dubins, 1962). Therefore, in the analysis that follows, we use the running mean[5] of the targets as a proxy for the participant's beliefs of the mean of the distribution of targets.

We note that sets of recent targets are simply noisy versions of the running mean. As such, tests involving averaged data will not be able to distinguish between the hypothesis that there is a bias toward the running mean and the hypothesis that there is a bias toward a set of recent targets. Unfortunately, HHV only analyze averaged data and therefore these two hypotheses are not distinguishable.[6] In this paper, we explore the extent to which the data can be explained by this alternate hypothesis.

Since the authors could not locate their datasets, we replicated the conditions in Experiment 3 from HHV. In our data, we find strong evidence of a bias toward recent targets and not toward the running mean of the distribution. This result is not consistent with CAM. In order to address the concern that our methods might not be able to detect a bias toward the running mean, we simulate data that exhibits a bias toward the running mean and not toward recent targets. Our methods correctly identify a bias toward the running mean and not toward recent targets in this simulated data.

Further, CAM is a mathematical model and this allows the researcher to devise non-obvious predictions that are consistent with the model. As Bayesian learners will have beliefs

[5] The average of the lengths of lines from the previous trials.
[6] The dangers of analyzing averaged data have been known in the psychology literature for some time (Sidman, 1952; Hayes, 1953; Estes, 1956; Siegler, 1987) and such concerns even appear in the recent judgments literature (Cassey, Hawkins, Donkin, & Brown, 2016; Hemmer, Tauber, & Steyvers, 2015).

that converge to the true distribution, if participants are Bayesian then we should observe evidence of learning across trials.[7] We subject the data to several tests of learning. We do not find evidence of the joint hypothesis that participants learn the distribution and employ this information in their judgments. These results are not consistent with CAM. Further, since there is no evidence of learning, it is difficult to see how this is consistent with any Bayesian model of judgment.

The entire contents of publications are assumed to be correct unless stated otherwise. Further, science can only be self-correcting if mistakes are identified and suggestions are made in order to avoid such mistakes in the future. A corollary to this is that declining to mention errors serves to prevent the progress of science. We therefore point out several problems with HHV, including dividing by zero. The alternative is to have incorrect aspects of a publication that are incorrect and assumed to be correct.

To our knowledge, Duffy and Smith (2018) is the only other paper on CAM to use the methods that we employ here. Duffy, Huttenlocher, Hedges, and Crawford (2010) claim that the results of their experiments are consistent with CAM. Duffy and Smith (2018) reexamine the data from Duffy et al. (2010) by analyzing judgment-level data rather than analyzing averaged data. Duffy and Smith do not find evidence of CAM in the Duffy et al. (2010) data. As we do here, Duffy and Smith find that there is a bias toward recent stimuli rather than toward the running mean of the distribution. Duffy and Smith also test whether there is evidence of learning

[7] It seems that the Bayesian judgment literature is unaware of the insights of Savage (1954) and Blackwell and Dubins (1962). However, these references have been in the psychology literature since Edwards, Lindman, and Savage (1963). On page 201, the authors state, “From a practical point of view, then, the untrammeled subjectivity of opinion about a parameter ceases to apply as soon as much data become available. More generally, two people with widely divergent prior opinions but reasonably open minds will be forced into arbitrarily close agreement about future observations by a sufficient amount of data. An advanced mathematical expression of this phenomenon is in Blackwell and Dubins (1962).”

the across trials and they fail to find evidence of learning. Duffy and Smith conclude that the Duffy et al. judgments are not consistent with CAM.[8] Here we perform an analysis similar to that of Duffy and Smith (2018).

The contributions of our paper are as follows. Contrary to the conclusions of HHV, we do not find evidence that the judgments are consistent with CAM. We do not find evidence of a bias toward the running mean and we also do not find evidence of the joint hypothesis that participants learned the distribution and employed this information in their judgments. It seems that evidence for CAM is a statistical illusion that appears when researchers analyze data averaged across trials and do not consider a recency bias. We also hope that our efforts lead to improved statistical techniques, including running multiple specifications, before arriving at strong conclusions. Additionally, it is our hope that the Bayesian judgment literature begins to include the insights from Savage (1954) and Blackwell and Dubins (1962) in the analysis of Bayesian models of judgment, including CAM. We point out specific technical problems with CAM in order to illustrate both the fundamental flaws of HHV and to explain the apparent lack of scrutiny that the paper received. Finally, our paper illustrates the importance of saving and sharing datasets.

We note that the *Journal of Experimental Psychology: General* (the outlet for HHV) declined to publish a previous version of this paper. Papers in print are assumed to accurate unless stated otherwise. Therefore, readers will conclude that HHV is entirely correct. It is therefore disappointing to us that the journal did not remedy any subset of the serious problems that we describe. A contribution of this paper is the description of the numerous and fundamental flaws – including dividing by zero – that continue to exist in the pages of a top psychology

[8] Also see Crawford (2019) and Duffy and Smith (2019).

journal. We hope that our efforts will lead to more forthcoming behavior from journals in admitting and correcting their flawed publications.

In Section 2, we describe CAM and its impact. In Section 3, we describe our replication of the conditions in Experiment 3 from HHV and in Section 4 we analyze the data. In Section 5, we discuss perhaps the most egregious mathematical errors in CAM. Section 6 concludes.

2. Category Adjustment Model

CAM purports that participants combine their noisy perception and memory of the target with their priors of the distribution of the targets. HHV offers (p. 239) the following formalism that response is a weighted average of the mean of the noisy, inexact memory of the target (M) and "the central value of the category" (ρ):

$$\text{Response} = \lambda M + (1-\lambda)\rho.$$

It is thus apparent that CAM is a model that generates the central tendency bias. The error in the memory of the targets is assumed to be normally distributed with a mean of zero. CAM therefore predicts that the mean response for the mean target will coincide with the target.

Further, CAM holds that the inexactness of the memory of the target has a standard deviation of σ_M and the "standard deviation of the prior distribution" is σ_P. The weight between M and ρ is a decreasing function g(.) of the ratio of these two standard deviations:

$$\lambda = g\left(\frac{\sigma_M}{\sigma_P}\right).$$

HHV describe another prediction of CAM as the following, "…the concentration of instances in the category should affect the variability of stimulus estimates. In particular, the variability of estimates of all categorized stimuli should be less when the prior distribution

(category) is more tightly clustered; this prediction, which follows from our Bayesian model, is not easily derived from other sets of assumptions" (p. 224).

Thus, CAM predicts that where the distribution of targets has more variance, responses will be closer to the target than to the mean of the distribution. We refer to these predictions of CAM as *explicit* because they were mentioned in HHV.

However, given that CAM is a model, we can also derive other predictions. As stated above, the smaller the standard deviation of the prior distribution, the greater the bias toward the mean of the distribution. We note that this decrease in standard deviation is precisely what happens over the course of an experiment. Before the participant has been exposed to any targets, the distribution is unknown and the participant relies on presumably diffuse priors. However, as the participant repeatedly views targets of various lengths, the standard deviation of the posteriors decreases across trials. The target lengths that have been observed will have increased posteriors across trials and the target lengths that have not been seen will have reduced posteriors across trials. This produces a decreasing standard deviation of the prior distribution across trials. Based on this, CAM predicts that the bias toward the running mean will increase over the course of the experiment.

Another such test of CAM is that, as participants observe the distribution across trials, Bayesian participants will improve their understanding of the distribution across trials. In fact, under mild assumptions, two Bayesian observers with different initial priors will both have posteriors that converge to the true distribution (Savage, 1954; Blackwell & Dubins, 1962).

Therefore, if participants make judgments consistent with CAM then they should learn the lower bound of the distribution and the upper bound of the distribution. In other words, participants should learn where the distribution has zero mass. If participants are Bayesian then

they should have diminishing priors across trials on line lengths that are longer than the maximum in the distribution and shorter than the minimum in the distribution. Accordingly, participants should offer such a response with a diminishing frequency across trials.

An additional implication of CAM relates to the errors across trials. HHV reports a monotonic relationship between the variance of the responses and the standard deviation of the prior distribution (σ_P). Therefore, as participants learn the distribution, errors should be decreasing across trials.

Below, we test these implicit predictions of CAM: whether responses with a zero mass in the distribution are declining across trials, whether there is an increased bias toward the running mean across trials, and whether errors are decreasing across trials.

Our analysis should be considered in light of the fact that CAM has had a very large impact on the literature. For example, CAM has been applied to topics such as the perception of neighborhood disorder (Sampson & Raudenbush, 2004), speech recognition (Norris & McQueen, 2008), overconfidence (Moore & Healy, 2008), categories of sound (Feldman, Griffiths, & Morgan, 2009), spatial categories (Spencer & Hund, 2002), spatial recall (Schutte & Spencer, 2009; Spencer & Hund, 2003; Hund & Spencer, 2003; Crawford & Duffy, 2010; Holden, Curby, Newcombe, & Shipley, 2010), visual illusions (Crawford, Huttenlocher, & Engebretson, 2000), delayed comparison of magnitude (Ashourian & Loewenstein, 2011), judgments of color (Bae, Olkkonen, Allred, & Flombaum, 2015; Olkkonen & Allred, 2014; Olkkonen, McCarthy, & Allred, 2014; Persaud & Hemmer, 2014), judgments of the size of familiar objects (Hemmer & Steyvers, 2009a, 2009b), judgments of the heights of people (Twedt, Crawford, & Proffitt, 2015), judgments of likelihood (Hertwig, Pachur, & Kurzenhäuser, 2005), facial recognition (Corneille, Huart, Becquart, & Brédart, 2004; Roberson, Damjanovic,

& Pilling, 2007; Young, Hugenberg, Bernstein, & Sacco, 2009), judgments of facial expressions (McCullough & Emmorey, 2009; Fugate, 2013; Corbin, Crawford, & Vavra, 2017), the perception of drink flavor (Woods, Poliakoff, Lloyd, Dijksterhuis, & Thomas, 2010), and judgments across different domains (Petzschner, Glasauer, & Stephan, 2015).

3. Our Replication of the Conditions of Experiment 3 in HHV

It is difficult for us to know how to replicate the experimental conditions for judgments of the fatness of computer generated fish (Experiment 1). Among other difficulties, HHV describe (p. 227) the fish as having an eye, yet the image of the fish in Figure 5 does not have an eye. It is thus not apparent to us where the eye should be located or how the position of the eye relative to the body would vary with changes in the fatness of the fish.

It is also difficult for us to know how to replicate the experimental conditions for judgments of the greyness of shades of grey (Experiment 2). Apparently, the light conditions are very important in this experiment. HHV report (p. 230) that, "The room was dimly lit by a single tungsten bulb." However, the authors do not provide a measure of the ambient light on the computer screen or even a specification of the bulb. Further, when the authors discuss the various shades of grey, HHV state (p. 230) that, "The units represent a linear transformation of photometer readings taken directly from the computer screen, resulting in a scale of darkness." However, the authors do not specify the units that they measure (candela, lux, lumen, etc.) and do not specify which linear transformation they use to convert the measures to those reported.

By contrast, a replication of the conditions for judgments of the length of lines (Experiment 3) would seem relatively straightforward. Therefore, in our view, it is easier to replicate the conditions of Experiment 3 in HHV than either Experiments 1 or 2.

3.1 Description of Methods

The experiment was conducted with E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA). The sessions were performed on standard 20 inch (51cm) HP monitors. E-Prime imposed a resolution of 1024 pixels by 768 pixels.

There were 4 between-subject treatments in which participants viewed and reproduced a series of lines randomly drawn from different frequency distributions. In each treatment, participants completed 192 trials. In each trial, participants saw the target line on the screen for 2 seconds. The screen then went blank for 1.2 seconds and an initial adjustable line appeared that was 8 pixels (0.35cm) in length. Participants manipulated the length of this line until they judged its length to be that of the target line. This was accomplished by using the "S" key (which made the line decrease in length) or the "L" key (which made the line increase in length).[9] Once satisfied that their adjustable line was equal to the length of the target line, participants pressed ENTER and the next trial commenced.

In the *short* treatment, participants viewed and reproduced target lines ranging in 16 pixel (0.7 cm) increments from 48 pixels (2.1 cm) to 224 pixels (9.8 cm) in length. Each of the 12 distinct line lengths was estimated once in 16 blocks. In the *long* treatment, participants viewed and reproduced targets ranging in 16 pixel increments from 240 pixels (10.5 cm) to 416 pixels (18.3 cm). Each of the 12 lengths was estimated once in 16 blocks. In the *uniform* treatment, participants viewed targets ranging in 16 pixel increments from 48 to 416 pixels. Each of the 24 lengths were estimated once in 8 blocks. Finally, in the *normal* treatment, participants viewed the

[9] Windows keyboard properties include 4 different repeat delay settings that range from "Long" (1) to "Short" (4). These experiments were conducted on setting 3. Windows keyboard properties also include 32 different repeat rate settings that range from "Slow" to "Fast." These experiments were conducted on the fastest setting. As these details are not reported in HHV, we do not know the corresponding conditions in the original experiment.

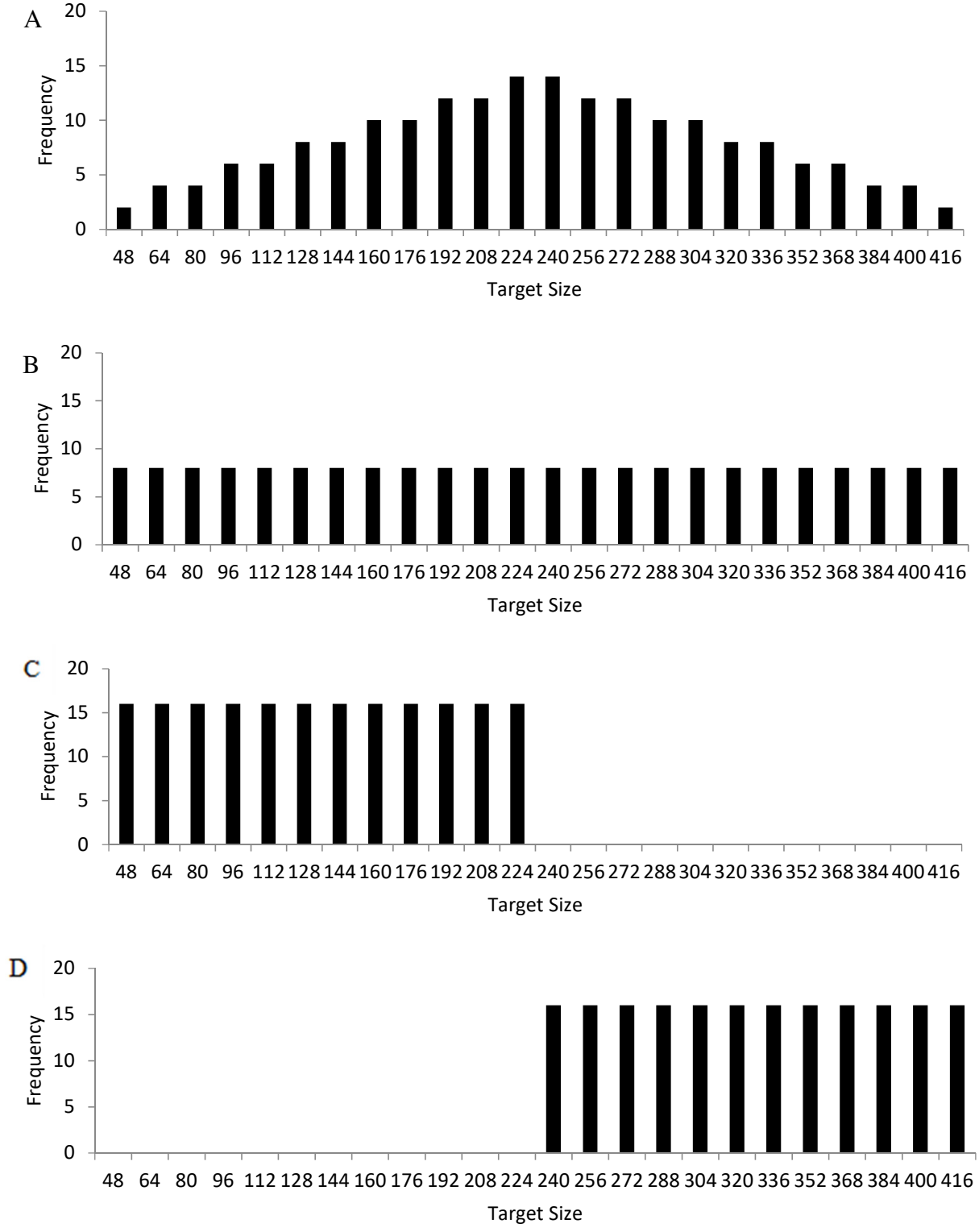
48 and 416 pixel lines once, the 64, 80, 384, and 400 pixel lines twice, the 96, 112, 352, and 368 pixel lines 3 times, the 128, 144, 320, and 336 pixel lines 4 times, the 160, 176, 228, and 304 pixel lines 5 times, the 192, 208, 256, and 272 lines 6 times, and the 224 and 240 lines 7 times. This constituted a single block. Upon completion, this block was repeated once more. See Figure 1 for a graph summarizing these distributions.

The thickness of each of these lines was 0.36 cm. In all four treatments, participants estimated a total of 192 lines, and there were no breaks between blocks.

The participants were given partial course credit for their participation.[10] There were 10 participants in the Normal treatment, 9 in the uniform treatment, 11 in the short treatment, and 11 in the long treatment.[11] With 41 participants each offering 192 judgments, we have a total of 7872 observations. We exclude 121 (1.54%) responses that are more than 3 standard deviations from the target.[12] This implies a total of 7751 observations. The study was approved by the Rutgers University Institutional Review Board.

[10] HHV offered a $5 show-up fee.
[11] HHV had 10 participants in each of the four treatments.
[12] On page 228, HHV describe their criterion, "…we calculated quartiles of the distribution of responses for each stimulus value, and we deleted responses deviating from the median by more than three interquartile ranges (IQRs)." On page 232, HHV report excluding "0.63% in the uniform condition, 0.36% in the normal, 0.63% in the short half, and 0.35% in the long half."

Figure 1: Target distribution for the normal (A), uniform (B), short (C), and long (D) treatments



3.2 A Discussion of the Design of Our Replication of the Experimental Conditions

We discuss the ways in which our attempted replication possibly deviates from Experiment 3 in HHV. One way in which this occurs is because the description of the design is not clear to us. For example, HHV report an inconsistent number of trials in the normal treatment. On page 232, the authors report numbers of trials within a block that sums to 106, which implies a total of 212 trials.[13] This contrasts with the other three treatments in Experiment 3 that have a total of 192 trials. In the description of Experiment 3, the HHV authors do not offer a justification for having treatments with unequal numbers of trials across treatments. In addition, the caption for Figure 4 in HHV, which provides histograms of the frequency distributions in all three experiments, suggests a number of trials that is different from that reported in the verbal description of Experiment 3. We decided that reporting 212 trials was likely an error in the paper and we designed the normal treatment to be consistent with those in Experiments 1 and 2, by having 192 trials.

On the other hand, some differences are due to the constraints imposed by our computer program. HHV has stimuli sizes that have an odd number of pixels, whereas we are constrained to have only an even number of pixels. As a result, HHV have lines that range in 15 pixel (0.5 cm) increments from 45 pixels (1.5 cm) to 390 pixels (13 cm). By contrast, our experiment has lines that range in 16 pixel increments from 48 pixels to 416 pixels.[14]

One additional difference relates to the initial adjustable line. HHV report an initial adjustable line of 2 pixels. However, this is shorter than the minimum line that we could produce on E-Prime. Our initial adjustable lines are 8 pixels.

---

[13] On page 232, HHV write, "In the normal conditions, the distribution of stimuli within each block was as follows: once at 45 and 390; twice at 60 and 375; three times at 75, 90, 345, and 360; four times at 105, 120, 315, and 330; five times at 135, 150, 285, and 300; six times at 165, 180, 255, and 270; and seven times at 195, 210, 225, and 240."

[14] We also note that the HHV lines had a thickness of 0.23 cm, rather than 0.36 cm in our experiment.

Finally, we used modern, 20 inch LCD screens whereas HHV used smaller CRT displays. The displays used in HHV are no longer commercially available.

We also concede that there are possibly differences between HHV Experiment 3 and our experimental conditions regarding the brightness of the display, the velocity with which the adjustable line increased or decreased in length, the distance between the participant and the display, etc. We also note that there could be differences between the sets of participants.

However, these differences only exist because the authors could not provide their datasets. Further, these experimental differences only affect the direct comparison of our data with those from HHV. These minor differences, on the other hand, do not diminish our data as offering a test of CAM. HHV does not claim that CAM only applies to the precise conditions of Experiment 3 but rather to judgments in general.

4. Results

4.1 Summary statistics

In order to compare our data with the data obtained by HHV, we look to their reported summary statistics. In this subsection, we restrict attention to one of the tools used by HHV: the t-test. In other words, the results in this subsection could have been reached by HHV using their techniques.

We define the *response bias* to be the response minus the target.[15] HHV report the standard deviations of the response bias in 6 different settings: the central 10 targets in the normal treatment, the central 10 targets in the uniform treatment, the short treatment, the shortest 12 targets in the uniform treatment, the long treatment, and the longest 12 targets in the uniform treatment. We list the standard deviations reported by HHV and the standard deviations in our

[15] HHV refer to this variable simply as *bias*. However, we also examine biases with different definitions, so we employ the term response bias.

data for the analogous treatments. As does HHV, we also test for the differences between the categories. HHV performs t-tests of the differences of the natural logs of the standard deviations.[16] We perform the identical analysis on our data. We report both the results of HHV and our results. We note that the reported natural logs of the standard deviations are calculated, not as the log of the average across targets, but as the average of the natural logs of the standard deviation within each target. We summarize this in Table 1.

Table 1: Standard deviations of response bias in HHV and our data

	SD_{HHV}	$\ln SD_{HHV}$	SD_{ours}	$\ln SD_{ours}$	Obs_{ours}	t_{HHV}	p_{HHV}	t_{ours}	p_{ours}
Normal, central 10	50.92	3.926	29.82	3.385	1155				
Uniform, central 10	60.27	4.083	34.75	3.544	711				
difference	-9.35	-0.157	-4.93	-0.159		-2.31	.033	-2.75	.013
Uniform, shortest 12	39.22	3.618	31.03	3.432	860				
Short	38.70	3.624	21.04	3.025	2105				
difference	0.52	-0.006	9.99	0.407		0.05	.960	6.18	<.001
Uniform, longest 12	77.66	4.345	40.27	3.692	828				
Long	52.31	3.948	39.88	3.683	2066				
difference	25.35	0.397	0.39	0.009		7.19	<.001	0.29	.78

Notes: We provide the standard deviations reported by HHV in Experiment 3 (Table 3) and the standard deviations in our data within the same setting. We report the average of the natural logs of the standard deviations for HHV and our data. We report the number of our observations within each category. We report the t-statistic of the difference between the natural logs and the p-value of a two tailed test, as reported in HHV and that for our data. The tests involving the normal and uniform distributions have 18 degrees of freedom. The remaining tests have 22 degrees of freedom. Although we note that HHV apparently incorrectly report 18 degrees of freedom for the short tests in their Tables 1, 2, and 3.

Similar to HHV, we find differences in two out of the three tests. The results are similar when we perform the tests on the raw standard deviations rather than the natural log of the standard deviations.[17]

[16] Given their reported degrees of freedom, it seems as if HHV conducted the tests assuming an equal variance between the samples. The reader might be concerned about the appropriateness of this. Our results are not changed when we conduct paired t-tests or unpaired t-tests that do not assume an equal variance.

[17] We find a significant difference between the normal and the uniform treatments ($t(18) = 2.75$, $p = .013$) and a significant difference between the short treatment and the shortest lines in the uniform treatment ($t(22) = 6.73$, $p <$

We also note that our participants were not less accurate than the HHV participants. We are therefore confident that the results that follow are not driven by excessively inattentive or inaccurate participants or by differences in keyboard speed.[18]

On their decision to restrict attention to the central 10 targets for the test of the difference between the normal and uniform treatments, HHV write, "We should restrict ourselves to a region within the categories where the certainty of membership is equal. Because the certainty that a stimulus is in the category decreases more markedly near the boundaries for a normal distribution than for a uniform distribution, we elected to compare standard deviations over a central region where participants were quite certain of the category for both distributions. Hence, we focused our attention on the 10 most central stimuli."[19]

We do not find this to be a compelling argument to exclusively examine the central 10 targets. We decided to investigate this matter by performing tests on a range of restricted target values. We perform a test on all of the data (24 targets), only the central 22 targets, only the central 20 targets, and so on, until only the central 6 targets. We perform these tests on both the raw data and the logged data. We summarize our analysis in Table 2.

.001). However, we do not find a significant difference between the long treatment and the longest lines in the uniform treatment ($t(22) = 0.29$, $p = .77$).

[18] We also note that we excluded 1.54% of trials whereas HHV excluded 0.49%.

[19] Page 229.

Table 2: Various t-tests for differences in standard deviations by target restrictions

Central	24	22	20	18	16	14	12	10	8	6
Normal	32.88	32.69	32.21	31.68	31.16	30.87	30.57	29.82	29.68	29.90
Uniform	35.65	35.53	35.46	35.34	35.03	34.85	34.64	34.75	34.61	34.88
t-statistic	-1.64	-1.61	-1.91	-2.31	-2.49	-2.34	-2.41	-2.75	-2.56	-2.61
p-value	.107	.114	.064	.027	.018	.027	.025	.013	.023	.026
ln Normal	3.477	3.470	3.457	3.443	3.428	3.418	3.409	3.385	3.382	3.394
ln Uniform	3.562	3.560	3.560	3.558	3.550	3.545	3.540	3.544	3.539	3.547
t-stat	-1.76	-1.75	-2.05	-2.39	-2.51	-2.39	-2.41	-2.75	-2.53	-2.63
p-value	.086	.088	.047	.022	.017	.024	.025	.013	.024	.025
Normal trials	1892	1858	1779	1701	1584	1469	1313	1155	958	758
Uniform trials	1688	1551	1414	1273	1131	991	853	711	572	430
% of total	100	95.2	89.2	83.1	75.8	68.7	60.5	52.1	42.6	33.2

Notes: We restrict attention to various central target lengths. For each, we list the average of the normal and uniform treatments, and the t-statistics and the p-values associated with a two-tailed test. The upper panel shows this for raw standard deviations and the middle panel for the natural log of the standard deviations. The lower panel reports the number of trials in the Normal and Uniform treatments considered and their percent of total.

The p-value attains its smallest value at the restriction to only the central 10 values. We further note that our p-values are .013 for both specifications in this restriction, whereas it is .033 for HHV. We do not know if the HHV data exhibit a similar relationship. We admit that 10 is a round number and this could have been the basis for the decision to report the test restricted to only the central 10 targets. However, it is curious that only a single specification[20] is reported by HHV (the central 10 targets) and, in our data, this happens to be restriction with the lowest p-value.

While HHV do not report the mean response bias, ours (M = -10.88, SD = 38.71) is significantly different from zero ($t(7750) = -24.74$, $p < .001$). Allred et al. (2016) found evidence that the length of the initial adjustable line affects judgments. We conjecture that these

[20] We use the term *specification* to refer to the complete set of assumptions in the analysis, including the functional form, the choice of explanatory variables, the assumptions regarding the error term, and the set of data under consideration.

underestimates, and the underestimates that we discuss below, are due to the short initial adjustable line. However, since the initial adjustable line is fixed, we are unable to test this conjecture.

Next we examine the mean of the response bias within treatments across targets. Figure 2, offers a summary of this data.

CAM predicts that each treatment will have a mean response bias of zero at the means of their distributions. However, this appears to not be the case in our data. We find that the mean response bias is negative in the uniform, normal, and long treatments.[21] When we restrict attention to the central two values in every treatment[22] we see similar results.[23] There seems to be a particularly stark difference in the response bias of the short and long treatments. Since the short treatment distribution and the long treatment distribution are identical (same number of targets, same frequencies, etc.) with the exception of the specific target sizes, this is a particularly troubling difference. We find that the mean response bias of judgments in the long treatment minus that of the short treatment is significantly different from zero, $t(22)= -3.99$, $p < .001$. This is robust to the specification of the test.[24] These significant relationships are clearly not consistent with CAM.

---

[21] Mean response bias is significantly less than zero in the normal treatment ($M = -10.40$, $SD = 38.09$, $t(1891) = -11.88$, $p < .001$), the uniform treatment ($M = -10.72$, $SD = 45.41$, $t(1687) = -9.69$, $p < .001$), and the long treatment ($M = -21.97$, $SD = 42.67$, $t(2065) = -23.41$, $p < .001$), but not in the short treatment ($M = -0.55$, $SD = 23.40$, $t(2104) = -1.08$, $p = .28$).

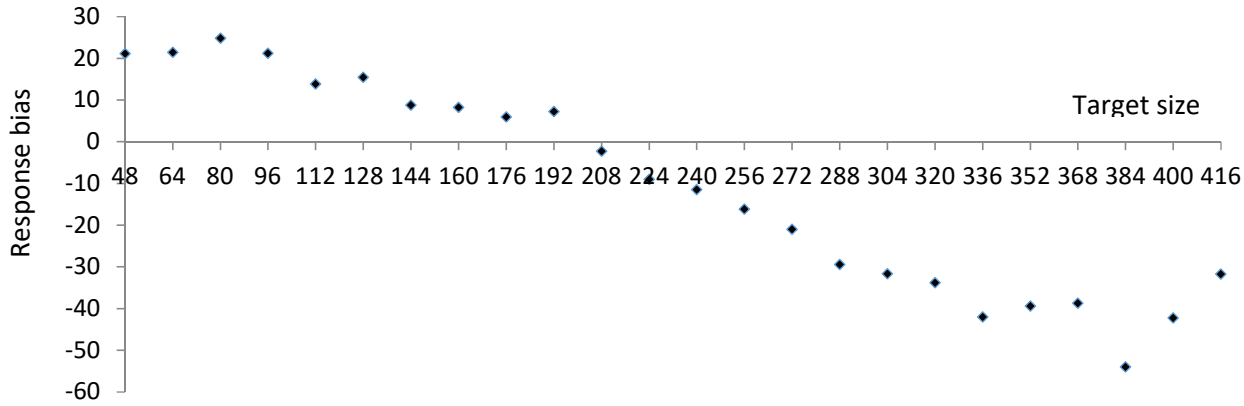
[22] Here we only include targets 224 and 240 in the normal and uniform treatments, targets 320 and 336 in the long treatment, and targets 128 and 144 in the short treatment.

[23] Restricted to the central two values, mean response bias is significantly less than zero in the normal treatment ($M = -10.27$, $SD = 27.65$, $t(278) = -6.20$, $p < .001$), the uniform treatment ($M = -9.61$, $SD = 32.89$, $t(144) = -3.51$, $p < .001$), and the long treatment ($M = -20.80$, $SD = 38.17$, $t(342) = -10.10$, $p < .001$), but not in the short treatment ($M = 1.31$, $SD = 21.73$, $t(351) = 1.13$, $p = .26$).

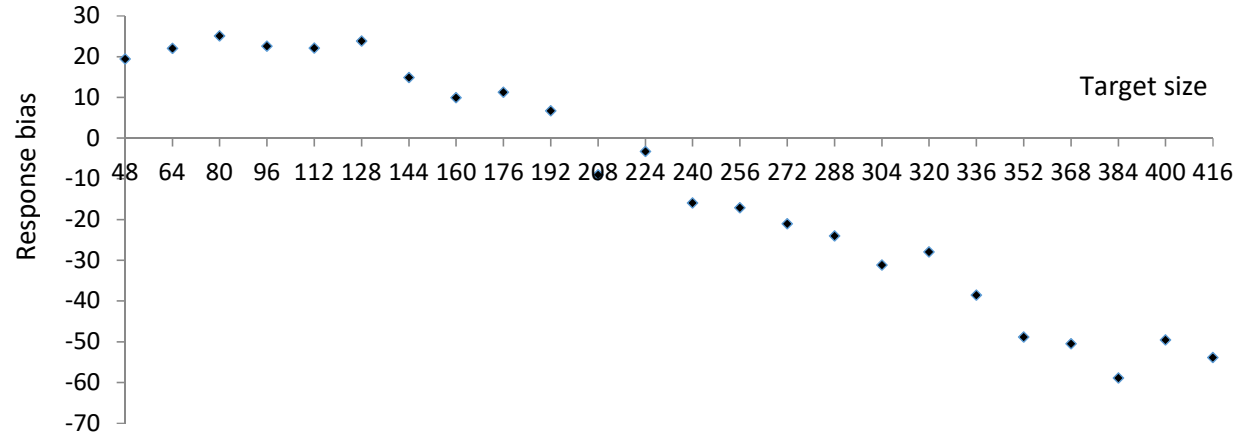
[24] We conduct a paired t-test ($t(11) = -10.85$, $p < .001$), an unpaired t-test that does not assume an equal variance ($t(18)= -3.99$, $p < .001$), and a t-test that does not assume an equal variance over all observations ($t(3191.3) = -20.05$, $p < .001$), and the results are not changed.

Figure 2: Response bias across targets for the normal (A), uniform (B), and short and long (C) treatments

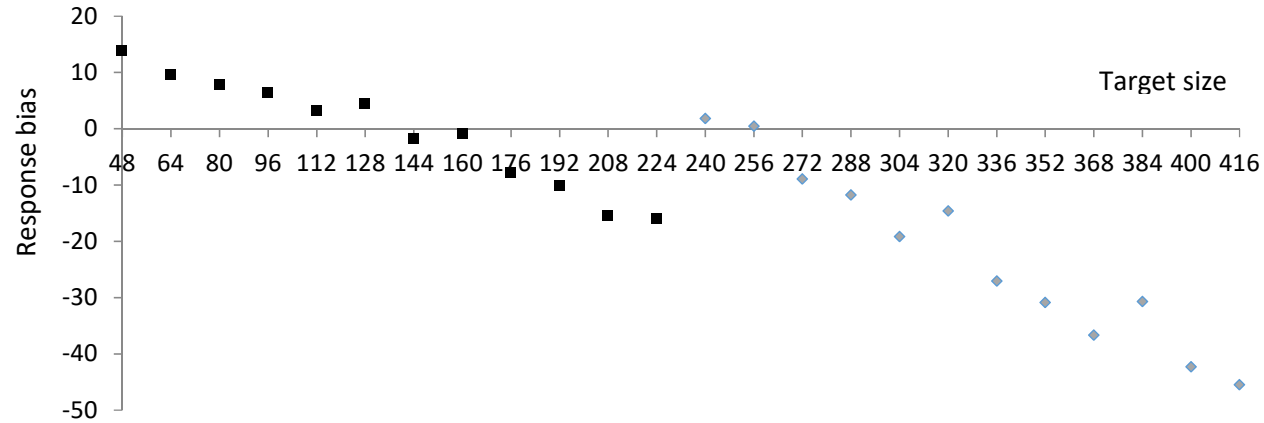
A



B



C



We do not know if these features exist in the HHV data, as the authors do not report any of these tests. Rather, the authors merely assert, "For all conditions, responses are shrunken toward a central value. There is overestimation of short line lengths and underestimation of long lengths" (p. 232). The first sentence is imprecise.[25] The second sentence is simply a restatement of the central tendency bias, which does not uniquely provide evidence in favor of CAM. It would have been preferable for HHV to report any subset of the tests that we report above. Below we will say more about the differences between the response biases in the short and long treatments.

4.2 Repeated measures regressions for running mean

CAM asserts that there is a bias toward the running mean of the stimulus sizes. Here we explore whether we find evidence of this. We define the *running mean* variable to be the mean of the targets that the participant has viewed in the previous trials. We conduct regressions with target and running mean as independent variables and response as the dependent variable.

In order to account for the lack of independence between two observations associated with the same participant, we employ a standard repeated measures technique. We assume a single correlation between any two observations involving a particular participant. However, we assume that observations involving two different participants are statistically independent. In other words we employ a repeated measures regression with a compound symmetry covariance matrix.[26]

[25] It is not clear to us why the authors were so vague or why the reviewers considered this to be evidence in support of CAM.

[26] We include the repeated measures because it is a better model. However, the results without repeated measures are qualitatively similar to those with repeated measures, in this and in subsequent analyses.

We restrict each of the regressions to a distribution treatment.[27] Since there is not a running mean on the first trial, we analyze data from trials 2 to 192.[28] These regressions are summarized in Table 3.[29]

Table 3: Random-effects repeated measures regressions of the response variable

	Normal	Uniform	Short	Long
Target	0.766***	0.753***	0.833***	0.730***
	(0.008)	(0.008)	(0.008)	(0.014)
Running mean	0.149*	0.083	-0.009	0.060
	(0.068)	(0.059)	(0.080)	(0.122)
-2 Log L	18290.1	16703.6	18746.8	20433.3
Observations	1882	1680	2095	2056

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts or the covariance parameters. † indicates significance at $p < .1$, * indicates significance at $p < .05$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

As would be expected, target is significantly related to response in every treatment. However, running mean is significant only in the normal treatment. In other words, in these specifications, without any other independent variable, there is only weak evidence of a bias toward the running mean.[30]

4.3 Repeated measures regressions for preceding target lines

Here we explore a simple alternate hypothesis to CAM: participants are affected by previous targets, rather than the running mean. We note that recency effects and sequential effects have been studied in the literature.[31] We perform an analysis similar to that summarized

[27] The pooled analysis appears in the "None" specification of Table 5.
[28] Regressions that analyze data from trials 2 through 192 have 7713 observations. The total 7751 minus 41 participants making judgments on the first trial, however 3 first trial judgments are excluded due to their inaccuracy.
[29] Table 3 and the regression tables that follow are not consistent with the American Psychological Association (APA) format for regressions. However, the APA format makes it difficult to display multiple specifications because the coefficient estimates and the standard errors are listed in separate columns. Since we prefer to display multiple specifications in each table, we present the regressions in a format, standard in other fields, with a regression in each column.
[30] This is robust to the specification of the error term. See Table A1 in the Supplemental Online Appendix. This is robust to a quadratic specification. See Table A10 in the Supplemental Online Appendix.
[31] See Jesteadt, Luce, and Green (1977), Staddon, King, and Lockhead (1980), Petzold (1981), Laming (1984), DeCarlo and Cross (1990), Choplin and Hummel (2002), Stewart, Brown, and Chater (2002), Petzold and

in Table 3 but we include an additional independent variable: the previous target. These regressions are summarized in Table 4.[32]

Table 4: Random-effects repeated measures regressions of the response variable

	Normal	Uniform	Short	Long
Target	0.766***	0.753***	0.835***	0.735***
	(0.008)	(0.008)	(0.008)	(0.014)
Running mean	0.104	0.051	-0.097	-0.022
	(0.069)	(0.060)	(0.080)	(0.123)
Previous target	0.030***	0.025**	0.053***	0.058***
	(0.008)	(0.008)	(0.008)	(0.014)
-2 Log L	18284.3	16701.3	18715.3	20423.0
Observations	1882	1680	2095	2056

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts or the covariance parameters. † indicates significance at $p < .1$, ** indicates significance at $p < .01$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

We note that running mean is not significant in any of the treatments. By contrast, previous target is significant at .01 in each treatment.[33] Finally, by comparing Tables 3 and 4, we note that the coefficient estimates for target are relatively unaffected by including previous target.

We also explore whether including additional sets of recently viewed stimuli can help predict response. As the analysis above, we include a specification that has an independent variable that is the preceding target line, which we refer to as *Prec 1*. We also calculate the average of the preceding 3, the preceding 5, and the preceding 10 target lines. We refer to these specifications, respectively, as *Prec 3*, *Prec 5*, and *Prec 10*. In order to maximize our data, Prec X is calculated as the mean of as many available previous targets as possible, but constrained to

Haubensak (2004), Wilder, Jones, and Mozer (2009), Yu and Cohen (2009), and Jones, Curran, Mozer, and Wilder (2013).

[32] The pooled analysis appears in the "Prec 1" specification of Table 5.

[33] This is robust to the specification of the error term. See Table A2 in the Supplemental Online Appendix. This is robust to a quadratic specification. See Table A11 in the Supplemental Online Appendix.

not be more than X. Our analysis below considers each of these 4 specifications for the preceding target line variables. We refer to this set of variables as *preceding targets*. We also include a specification without any information about the previous targets, which we label as *None*. Finally, because the results of Tables 3 and 4 suggest that there are differences among the treatments, we estimate a dummy variable for each treatment. These regressions are summarized in Table 5.

Table 5: Random-effects repeated measures regressions of the response variable

	None	Prec 1	Prec 3	Prec 5	Prec 10
Target	0.765***	0.766***	0.766***	0.766***	0.766***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Running mean	0.0870*	0.0383	0.0274	0.0438	0.0293
	(0.0368)	(0.0372)	(0.0384)	(0.0398)	(0.0435)
Preceding targets	-	0.0343***	0.0444***	0.0331**	0.0478*
		(0.0045)	(0.0084)	(0.0117)	(0.0193)
-2 Log L	74784.3	74734.9	74763.9	74783.3	74784.2

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts, the treatment dummies, or the covariance parameters. All regressions have 7713 observations. † indicates significance at $p < .1$, * indicates significance at $p < .05$, ** indicates significance at $p < .01$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

In the specification without any information about previous lines, running mean is significant. However, in each of the specifications that include information about previous stimuli, running mean is not significant, whereas preceding targets are significant. This suggests that the preceding lines are much better predictors of responses than the running mean.[34]

4.4 Analysis of simulated data consistent with a key feature of CAM

Given the stark results above, a researcher might be concerned that that our techniques are not sufficiently sensitive to detect evidence of CAM. In particular, a researcher might note that the standard deviation of running mean decreases across trials and this might prevent a

[34] This is robust to the specification of the error term. See Table A3 in the Supplemental Online Appendix. This is robust to quadratic specifications. See Table A12, A13, and A14 in the Supplemental Online Appendix.

satisfactory inference of the running mean coefficient. In order to investigate this matter, we simulated a simple dataset that is consistent with a key feature CAM and has parameters similar to those found in our data. We took the sequence of targets and added to each a normally distributed noise term, with a zero mean and a standard deviation of 25 pixels. We refer to the sum of target and the noise as the *memory* variable. We then define the *response25* variable to be the weighted average of memory and running mean. Although our analysis above suggests that an increase of the running mean by 1 pixel would lead to a larger response by .087 pixels, here we use a weight of .08:

$$\text{Response25} = .92(\text{Memory}) + .08(\text{Running mean}).$$

These simulated judgments are consistent with a key feature of CAM in that response25 is biased toward running mean but not toward recent lines. Additionally, there is a slightly lower weight on running mean than in the original dataset. Therefore, detecting a relationship between running mean and response is slightly more difficult in our simulated data than in the original data. We perform the identical analysis to that performed in Table 5, which we summarize in Table 6.

Table 6: Random-effects repeated measures regressions of the simulated response25 variable

	No Prec	Prec 1	Prec 3	Prec 5	Prec 10
Target	0.921***	0.921***	0.921***	0.921***	0.921***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Running mean	0.103***	0.102***	0.104***	0.116***	0.107***
	(0.026)	(0.027)	(0.028)	(0.028)	(0.031)
Preceding targets	-	0.0009	-0.0006	-0.0103	-0.0037
		(0.0033)	(0.0062)	(0.0086)	(0.0142)
-2 Log L	71295.1	71304.6	71303.4	71301.4	71301.7

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts, the treatment dummies, or the covariance parameters. All regressions have 7831 observations. † indicates significance at $p <$ .1 and *** indicates significance at $p <$ .001. -2 Log L refers to negative two times the log-likelihood.

In every specification, running mean is significant at .001 and preceding targets is not significant. In the Supplemental Online Appendix, we include an analysis similar to Table 6,

performed on the response35 variable, which has a standard deviation of 35 not 25. The qualitative results hold.[35,36]

Why are there such stark differences between the results from the simulated and non-simulated data? The mean of a set of recent lines is a noisy version of the running mean. But unless there is actually a bias toward recent lines, recent lines will be worse predictors than the running mean. Our simulated data does not have a bias toward recent lines. Accordingly, our analysis does not detect such a bias. On the other hand, in our non-simulated data there is a bias toward the running mean and our analysis identifies this to be the case. In summary, we reject the claim that our techniques are unable to detect a bias toward the running mean, should such a bias exist.

For the reader worried about multicollinearity driving the results in Table 5 (and Table A3), we note that the identical multicollinearity exists in Table 6 (and Table A4). However, with the simulated data we find the relationship predicted by CAM but in the non-simulated data we do not find such a relationship. Therefore, we reject this potential objection.

4.5 Responses with zero mass across trials

Although our tests of the explicit predictions of CAM fail to find evidence in support of the model, we now look for evidence of learning across trials. If participants are Bayesian then they should have diminishing priors across trials on line lengths that are longer than the maximum in the distribution or shorter than the minimum in the distribution. Accordingly, participants should offer such a response with a diminishing frequency across trials. One such

[35] See Table A4 in the Supplemental Online Appendix.
[36] For the reader worried about multicollinearity driving the results in Table 5 (and Table A3), we note that the identical multicollinearity exists in Table 6 (and Table A4). However, with the simulated data we find the relationship predicted by CAM but in the non-simulated data we do not find such a relationship. Therefore, we reject this potential objection.

test of CAM is that, there should be a declining incidence of responses that are shorter than the minimum target or longer than the maximum target.

We define the *zero mass dummy* to be 1 if the response is greater than the maximum in the distribution[37] or less than the minimum of the distribution,[38] and a 0 otherwise. In Figure 3 we plot the average of this variable across trials.

Figure 3: Mean of the zero mass dummy variable across trials

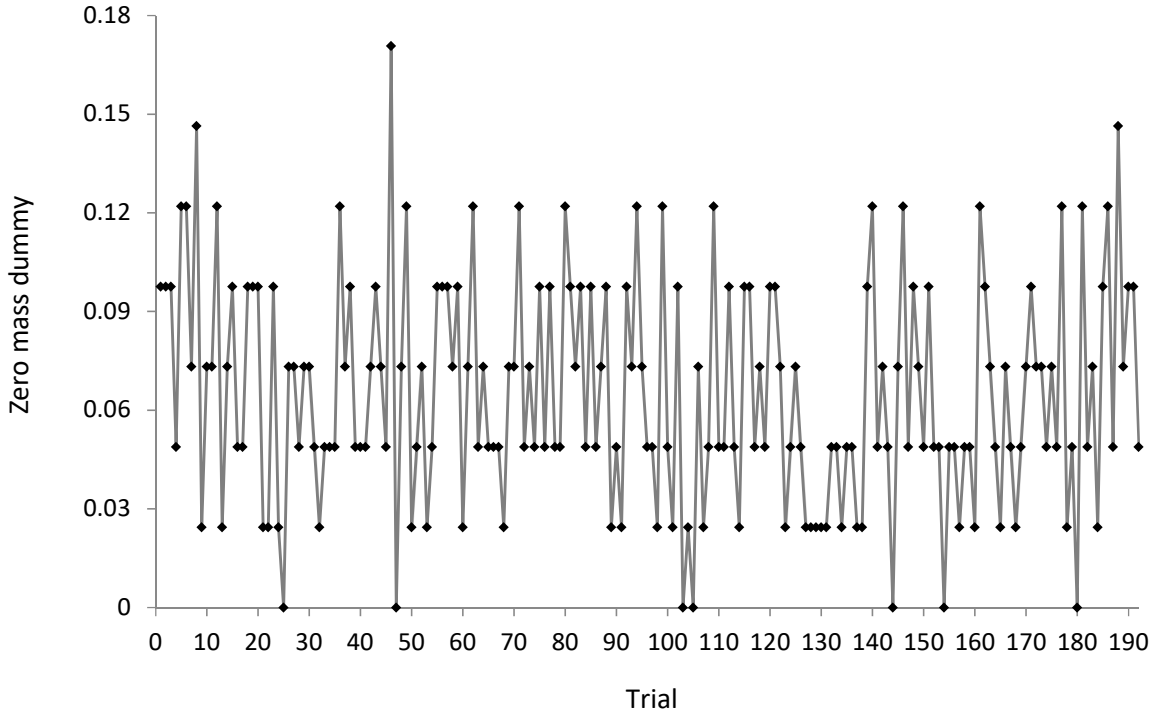


Figure 3 suggests that the zero mass dummy is not decreasing across trials. To test this, we perform the following analysis. We offer different measures of the rate of the learning. In one specification, the independent variable is simply the trial number. But perhaps the learning is not linear and follows the square root. Therefore, we offer a second specification where the independent variable is the square root of the trial number, which we refer to as *Sqrt. Trial*. In

[37] For the uniform, normal, and long treatments the maximum is 416. For the short treatment the maximum is 224.
[38] For the uniform, normal, and short treatments the minimum is 48. For the long treatment the minimum is 240.

the remaining four specifications, we use a categorical variable indicating whether the trial is among the first 5, among the first 10, among the first 20, or among the first half of trials.

We conduct the analysis similar to those above but with some differences. First, due to the discrete nature of the zero mass dummy, we conduct a logistic regression. Second, we account for the repeated measures by a fixed-effects regression. In other words, we estimate a unique dummy variable for every participant. Third, the zero mass dummy might depend on the target size and the treatment, so we control for this possibility by estimating a dummy variable for each target in each distribution treatment. Table 7 summarizes this fixed-effects analysis. We note that CAM predicts negative estimates for Trial and Sqrt. Trial, but positive estimates for the others. There are 455 responses with a zero mass and 7296 without.

Table 7: Fixed-effects logistic regressions of the zero mass dummy variable

	Trial	Sqrt. Trial	First 5	First 10	First 20	First half
Trial	-0.0015	-0.026	0.062	0.199	0.283†	0.251*
	(0.001)	(0.017)	(0.355)	(0.250)	(0.171)	(0.111)
-2 Log L	2247.2	2247.0	2249.3	2248.7	2246.7	2244.2

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 1 through 192. We do not provide the estimates of the intercepts, the participant dummy variables, or the treatment-target dummy variables. All regressions have 7751 observations. † indicates significance at $p < .1$ and * indicates significance at $p < .05$. -2 Log L refers to negative two times the log-likelihood.

We find evidence of learning but only in the First half specification. The reader might be concerned that participants exhibit exhaustion over the entire 192 trials. Accordingly we analyze only the first half of the trials.[39] There we do not find evidence of learning as measured by the zero mass dummy variable. Tables 7 and A5 produce a total of 11 specifications and we find a significant relationship in only one specification. We also note that in the 11 specifications, three do not even have the correct sign as predicted by CAM.

4.6 Bias toward the running mean across trials

[39] See Table A5 in the Supplemental Online Appendix.

Another indirect test relates to the bias toward the running mean across trials. We construct a variable that is designed to capture the extent to which the response is closer to the mean than it is to the target. We define *running mean bias* to be the distance between the target and the running mean minus the distance between the response and the running mean:

Running mean bias = | Target – Running mean | – | Response – Running mean |.

We perform a random-effects repeated measures analysis, similar to that summarized in Tables 3-6. However, we employ the independent variables in the analyses summarized in Table 7. Table 8 summarizes this random-effects analysis. CAM predicts positive estimates for Trial and Sqrt. Trial and negative estimates for the others.

Table 8: Random-effects regressions of the running mean bias variable

	Trial	Sqrt. Trial	First 5	First 10	First 20	First half
Trial	0.0251***	0.486***	-7.176**	-5.002***	-5.004***	-2.277***
	(0.00556)	(0.0960)	(2.178)	(1.463)	(1.026)	(0.613)
-2 Log L	72401.9	72145.3	72399.4	72399.4	72388.1	72399.0

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 7713 observations. ** indicates significance at $p < .01$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

Here we find strong evidence of an increase in the bias toward the running mean across trials. We also note that this result is robust to restricting attention to the first half of trials.[40] It is further robust to expressing the bias toward the running mean as a ratio, as this measure would be more similar to the weight (λ) between the memory and the distribution in CAM. With running mean bias ratio, we also find strong evidence of an increasing bias toward the running mean across trials.[41,42]

[40] See Table A6 in the Supplemental Online Appendix.
[41] See Table A7 in the Supplemental Online Appendix.
[42] We note that most of the results that we report in this paper are similar to those reported in Duffy and Smith (2018). The exception is the relationship between running mean bias and trials. Duffy and Smith do not find a relationship in the Duffy et al. (2010) data, but here we find a strong relationship.

These results seem to be consistent with CAM but a deeper look into these results suggests otherwise. We perform an analysis to learn whether there is an increasing bias toward the previous line across trials. We therefore construct the variable *previous bias*, which is analogous to running mean bias. We conduct the analysis identical to that in Table 8 but with this new dependent variable. This analysis is summarized in Table 9.

Table 9: Random-effects regressions of the previous bias variable

	Trial	Sqrt. Trial	First 5	First 10	First 20	First half
Trial	0.0220**	0.447***	-9.920***	-7.090***	-5.000***	-1.977**
	(0.0069)	(0.118)	(2.679)	(1.800)	(1.264)	(0.754)
-2 Log L	75551.2	75541.6	75535.9	75534.9	75535.4	75545.2

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 7713 observations. ** indicates significance at $p < .01$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

In every specification, there is a greater bias toward the previous line across trials. Whereas the results of Table 8 are consistent with both learning the distribution and employing this information in judgments, the results in Table 9 are not consistent with this explanation. When we restrict attention to the first half of trials we also see strong evidence of an increase in the bias toward the previous lines across trials.[43,44]

4.7 Error across trials

An additional indirect prediction of CAM is that, as participants learn the distribution, the errors will diminish across trials. We define the *absolute response bias* variable to be the absolute value of the response bias. Absolute response bias is a measure of the error of the judgment. We perform an analysis similar to Table 9, but with absolute response bias as the

[43] See Table A8 in the Supplemental Online Appendix.

[44] Interestingly, we note a positive correlation between the response time and previous bias ($r(7711) = .038$, $p = .002$) but no such relationship between response time and running mean bias ($r(7711) = -.015$, $p = .19$).

dependent variable. Table 10 summarizes this random-effects analysis. CAM predicts negative estimates for Trial and Sqrt. Trial, and positive estimates for the other specifications.

Table 10: Random-effects regressions of the absolute response bias variable

	Trial	Sqrt. Trial	First 5	First 10	First 20	First half
Trial	0.0325***	0.568***	-2.016	-3.777**	-4.542***	-2.328***
	(0.0047)	(0.081)	(1.679)	(1.195)	(0.860)	(0.524)
-2 Log L	70390.9	70383.2	70424.9	70417.0	70399.8	70408.9

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 1 through 192. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 7751 observations. † indicates significance at $p < .1$, ** indicates significance at $p < .01$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

Rather than errors diminishing across trials, we see errors increasing in the Trial, Sqrt. Trial, First 10, First 20, and First half specifications. These results are not consistent with participants learning the distribution and using this knowledge to improve their judgments. Perhaps exhaustion is driving these results. Therefore, we analyze only the first half of trials and our results are similar to that from Table 10.[45] In conclusion, not only do we not find evidence that judgments improve across the trials, we observe that judgments become worse across trials. In other words, even though running mean bias increases across trials, this does not appear to be consistent with learning the distribution, which is a central component to CAM.

4.8 Difference between the long and short treatments across trials

Earlier we found that there were differences in the response bias between the long and short treatments, and that this is not consistent with CAM. On the other hand, it might not be reasonable to expect that these differences diminish before participants learned the distribution. According to this view, we should see the difference in the response bias in these treatments converging to zero as participants learn the distribution. In Figure 4, we plot the average response bias across trials, for the long and short treatments.

[45] See Table A9 in the Supplemental Online Appendix.

Figure 4: Mean response bias for the long and short treatments across trials

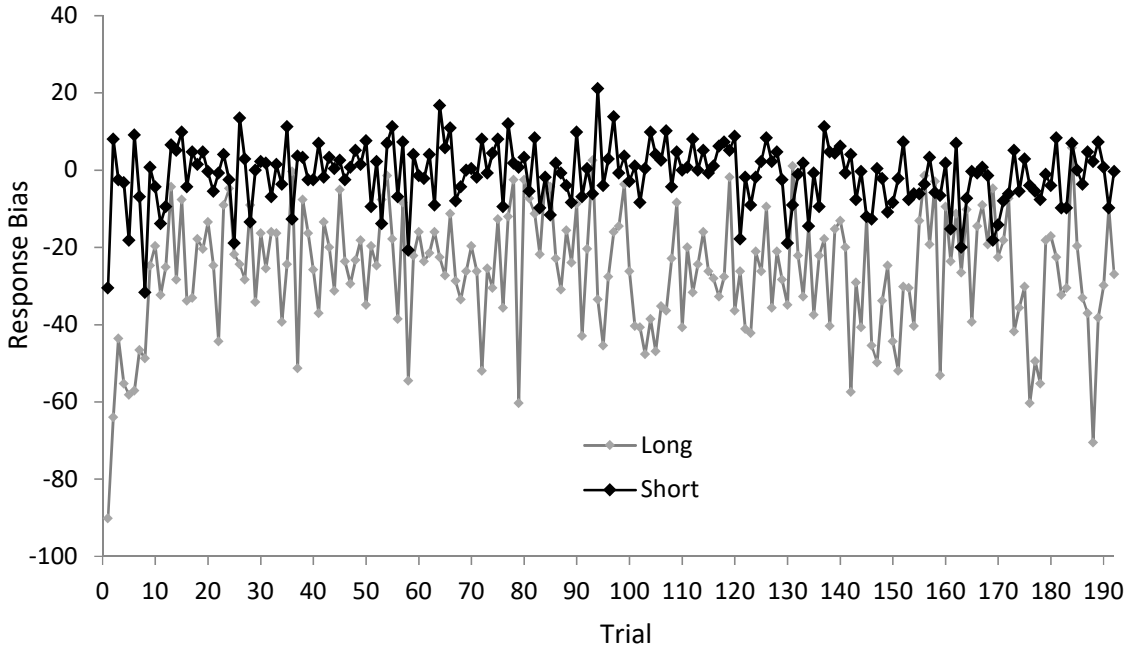


Figure 4 suggests that the difference between the response bias in the long treatment and the short treatment does not converge to zero.

We also acknowledge that Figure 4 is rather noisy. Further, from Tables 3-5 we know that response bias varies by the target. Therefore, in the analysis below, we control for the target. We define the *normalized target* to be the target minus the mean of the distribution. In the short treatment, the normalized target is the target minus 136, and in the long treatment, the normalized target is the target minus 328. This variable allows us to compare the shortest target in the short treatment with the shortest target in the long treatment, the 2nd shortest line in the short treatment with the 2nd shortest line in the long treatment, and so on. We also employ the variety of independent variables for the trials, as was used in the analyses summarized in Tables 7-10. We define *short* to be a dummy variable that indicates whether the trial is the short treatment. We also include the interactions with the relevant measures of trials. If participants are

learning the distribution and employing this information then we would see these differences declining across trials. This analysis is summarized in Table 11.

Table 11: Random-effects regressions of the response bias variable

	Trial	Sqrt. Trial	First 5	First 10	First 20	First half
Intercept	-22.28***	-22.85***	-21.80***	-21.80***	-21.78***	-22.60***
	(4.79)	(5.00)	(4.66)	(4.66)	(4.67)	(4.71)
Norm. target	-0.220***	-0.220***	-0.219***	-0.220***	-0.220***	-0.220***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Short	23.05**	23.67**	21.18**	21.35**	21.24**	21.32**
	(6.77)	(7.07)	(6.59)	(6.59)	(6.60)	(6.65)
Trial	0.0033	0.096	-6.602	-3.289	-1.755	1.277
	(0.012)	(0.196)	(4.111)	(2.930)	(2.085)	(1.262)
Trial*Short	-0.018	-0.250	7.187	0.145	1.067	0.072
	(0.016)	(0.276)	(5.728)	(4.085)	(2.928)	(1.776)
-2 Log L	39928.6	39918.0	39904.2	39905.7	39908.6	39909.3

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 1 through 192. We do not provide the estimates of the covariance parameters. All regressions are restricted to the short or the long treatments and have 4171 observations. † indicates significance at $p < .1$, ** indicates significance at $p < .01$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

We interpret the intercept as the average response bias for the mean target in the long treatment when the independent variables are zero, and trial as describing the trajectory of the long treatment estimate across trials. In none of the specifications do we see that the average response bias in the long treatment increases toward zero. We interpret the short variable as an estimate of the amount that judgments in the short treatment is larger than those in the long treatment at the mean of the targets and when the other independent variables are zero. We also interpret the interaction of short and trial as difference in the trajectory of the short and long treatments. We also do not find evidence of learning here.

To confirm what we suspected from Figure 4, we do not find evidence that the difference in the average response bias is diminishing across trials. There appears to be persistent differences between the treatments. Again, this is not consistent with CAM.

Since the length of the initial adjustable line does not vary across trials, we cannot distinguish between the hypothesis that it is caused by the short initial adjustable line or the hypothesis that that there is a bias toward the center of the screen. But regardless of the cause, these results are not consistent with CAM.

5. A few comments on the mathematical content of HHV

Figure 3 in HHV is justified by the mathematical content on page 241. We turn our attention to the relevant text: Section Vll, Subsection A. There HHV considers the relationship between the variance of the responses to the standard deviation of the distribution. The authors write the variance as:

$$S(R) = g\left(\frac{\sigma_M}{\sigma_P}\right)\sigma_M,$$

and the partial derivative with respect to σ_P as:[46]

$$S'(R) = -\left(\frac{\sigma_M}{{\sigma_P}^2}\right) g\left(\frac{\sigma_M}{\sigma_P}\right).$$

HHV define g(0)=1 and g(c)=0 for some "large" c. Apparently g(x) is not defined for x > c. The authors investigate the implications of σ_P=0. But this has the undesirable feature of dividing by zero, once in S(R) and twice in S'(R). Further, the authors cannot be making an argument about the limits as σ_P converges to zero because, for any c there exists a σ_P>0 small enough so that g(.) is not defined.

It is troubling to see an expression that divides by zero and considers values that are not defined. Further, given that the experiment only considers settings where $\sigma_P > \sigma_M$ and that HHV does not suggest how to generate $\sigma_P = 0$ in the laboratory, it is our view that restricting attention to $\sigma_P > \sigma_M$ is reasonable.

[46] We use the notation of HHV. However, it is not clear to us why the authors chose to employ a non-standard notation for the partial derivatives. Standard notation would be $\frac{\partial S(R)}{\partial \sigma_P}$.

Next we turn our attention to Subsection B. There HHV considers the relationship between the response bias and the standard deviation of the prior distribution. The authors represent the response bias as:

$$B(R) = \left(g\left(\frac{\sigma_M}{\sigma_P}\right) - 1\right)(\mu - \rho),$$

where μ is the mean of the noisy memory of the line length and ρ is the "central value of the distribution." The authors state that partial derivative with respect to σ_P is "difficult to compute," because "σ_P depends on μ." First, it is not at all clear to us why the standard deviation of the prior distribution should depend on the mean of the noisy memory of a particular target line length under consideration. Accordingly, we calculate the partial derivative of the response bias with respect to σ_P as:[47]

$$B'(R) = -\left(\frac{\sigma_M}{{\sigma_P}^2}\right) g'\left(\frac{\sigma_M}{\sigma_P}\right)(\mu - \rho).$$

Again the authors investigate the response bias at σ_P=0. Again, this constitutes dividing by zero. And further, arguments on the limit as σ_P converges to zero do not make sense because there are values of $\sigma_P > 0$ small enough so that g(.) is not defined. The authors make non-monotonicity arguments that seem to crucially depend on dividing by zero. HHV write "B(R) is never monotonic as σ_P is varied because B(R) always has a maximum." We do not understand why the authors assert this as true and it seems to rely on arguments where g(.) is not defined. If we restrict attention to the setting of the experiment ($\sigma_P > \sigma_M$) then it would seem that we are left with a monotonic relationship between response bias and σ_P.

[47] Again, we use the notation of HHV. Standard notation would be $\frac{\partial B(R)}{\partial \sigma_P}$.

These problems go beyond simply identifying an incorrect spelling or a typo.[48] Making predictions from a model where the authors divide by zero is problematic. The problematic aspects of dividing by zero are not limited to the appendix. For instance, both panels in Figure 3 of HHV characterize situations where the denominator is zero.

6. Conclusions

Since the authors of HHV were not able to provide their data to us, we replicated the conditions from their Experiment 3. Our data and their data are similar in many respects. Both HHV and our results indicate that 2 out of 3 comparisons reported by HHV have significantly different standard deviations. We also note that our judgments are not less accurate than the HHV judgments. We further note that our data shares some qualitative features with HHV, for instance in each of the four treatments, there is a negative relationship between the response bias and the target length. However, in our data, we find that judgments in the normal, uniform, and long treatments have a mean response bias that is negative. This persists when we restrict the analysis to the targets adjacent to the means in the distributions. We also find that the 12 target lengths in the short treatment are overestimated relative to the 12 targets in the long treatment. This is not consistent with CAM. We do not know if these results exist in the HHV data because they do not report a test of these features.

Further, HHV analyzed data averaged across previous stimuli. This renders the hypothesis that there is a bias toward the running mean and the hypothesis that there is a bias toward recent targets to be indistinguishable. By contrast, we conduct an analysis of the judgment-level data in order to determine if there is a bias toward the running mean or a bias toward recent targets. Our analysis shows that there is not a bias toward the running mean but

[48] Although we note that HHV (page 233) refer to "dark" trials in the discussion of Experiment 3. Recall that Experiment 2 in HHV involved judgments of color shade but Experiment 3 involved judgments of length.

rather a bias toward recent targets. In order to address the concern that our techniques would not be able to detect a bias toward the running mean, should such a bias exist, we simulate data that has a bias toward the running mean but not toward recent lines. Our techniques correctly identify this relationship. We therefore reject the criticism that our techniques would not identify a bias toward the running mean, should such a bias exist in the data.

HHV also analyzed data averaged across trials, and therefore learning properties are not able to be examined. We test some implications of CAM related to participants learning the distribution of targets and employing this information in their judgments. We do not find evidence that responses that have a zero mass are declining across trials. While we find evidence of an increase of the running mean bias across trials we also find evidence of an increase in the previous bias across trials. Additionally, we find that the errors in the judgments increase, rather than decrease, across trials. Finally, we find that the difference in the response bias between the long and short treatments does not diminish across trials. In summary, we do not find evidence that participants are learning the distribution and are employing this information to improve their judgments. These results are not consistent with CAM.

Taken together we simply do not find evidence that the judgements in our data are consistent with CAM. In addition to Duffy and Smith (2018), this is now the second paper that examines a dataset from an experiment that was previously considered to be consistent with CAM, however careful judgment-level analysis shows that it is not consistent with CAM. Evidence for CAM seems to be a statistical illusion that appears when researchers analyze data averaged across trials and do not consider a recency bias.

More generally, CAM is a *Bayesian model of judgment*. Specifically, Bayesian models of judgment make the joint hypothesis that participants learn the distribution of stimuli and they use

this information in their judgments in accordance with Bayes' rule. There is a spirited discussion of the merits of these Bayesian models.[49] We contribute to this literature by demonstrating that a judgment-level analysis shows that our data are inconsistent with CAM. We encourage researchers to employ our techniques in different settings in order to learn the extent to which the predictions of CAM, or any Bayesian model of judgment, are accurate.

Further, aside from Duffy and Smith (2018), we are the first to apply to Bayesian models of judgment the well-known results that Bayesians with different initial priors will have posteriors that converge to the true distribution (Savage, 1954; Blackwell & Dubins, 1962). In our analysis we do not see evidence of learning, either because there was no learning or because the learning did not manifest itself in the judgments. Regardless, it does not seem that our results could be consistent with any Bayesian model of judgment.

One question is, "How could the shortcomings of HHV go unnoticed?" For instance, it is not clear to us how the analysis of HHV could *verify* that the bias in judgments does not stem from recent stimuli and that the judgments are Bayesian. It is our view that the mathematical content of HHV contributed to its lack of scrutiny. The inclusion of mathematical formalism, even if it is unrelated to the topic, enhances the perception of the quality of the research, particularly among those with less mathematical skill (Eriksson, 2012). It is possible that the mathematical content of HHV dissuaded readers and reviewers from carefully judging the paper. Paradoxically, this includes noticing the errors in the mathematics itself. [50] Everybody knows that one should not divide by zero. Yet, there on page 241 we find multiple instances of dividing by

[49] See Barth, Lesser, Taggart, and Slusser (2015), Bowers and Davis (2012a, 2012b), Cassey et al. (2016), Chater, Tenenbaum, and Yuille (2006), Chater et al. (2011), Duffy and Smith (2018), Elqayam and Evans (2011), Goodman et al. (2015), Griffiths, Chater, Norris, and Pouget (2012), Griffiths and Tenenbaum (2006), Hahn (2014), Jones and Love (2011a, 2011b), Hemmer and Steyvers (2009a, 2009b), Lewandowsky, Griffiths, and Kalish (2009), Marcus and Davis (2013, 2015), Mozer, Pashler, and Homaei (2008), Perfors, Tenenbaum, Griffiths, and Xu (2011), Petzschner, Glasauer, and Stephan (2015), Rahnev and Denison (2018), Sailor and Antoine (2005), Tauber et al. (2017), and Tenenbaum, Griffiths, and Kemp (2006).

[50] For instance, we note two violations of the chain rule on page 239.

zero. It is surprising that we are possibly the first researchers to notice this since HHV was submitted for review.

Further, consider the term "fine-grained memory" that is used to refer to the memory of the length of the particular stimulus. This term appears throughout HHV. It is not clear to us how this is an improvement over "stimulus memory" or simply "memory." This is particularly true since there are not comparisons of memories with more or less granularity. The use of this term is an example of, what in our view, is opaque language employed by HHV. A consequence of this opaque language is that the reader can suffer from the "Guru effect" (Sperber, 2010) whereby the reader confers more authority and plausibility to a paper when it contains opaque language. It is our view that the opaque writing of HHV also contributed to its lack of scrutiny.

Additionally, HHV always offered analyses with a single specification, which we admit is standard in the psychology literature. In other words, HHV uses only a single type of test, a single set of explanatory variables, a single functional form, a single set of assumptions for the error term, and a single set of data under consideration. This becomes all the more serious given the choice of examining only the central 10 values for the standard deviations of the normal and uniform treatments. Reporting more than one specification, as we make a point of doing, can help diminish the chances of arriving at incorrect conclusions and give the reader a greater confidence in the results (Simmons, Nelson, & Simonsohn, 2011; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). We hope that our efforts here contribute to the ongoing discussion of improving the methods and conventions of psychological science (Wagenmakers, Wetzels, Borsboom, & Maas, 2011; Wicherts et al., 2016).

Further, our efforts highlight the importance of maintaining and sharing datasets so that researchers can scrutinize their results. We note the evidence that datasets in the more distant past tend to be less available than more recent datasets (Vines et al., 2014).

Finally, after describing numerous, fundamental flaws in HHV, we restate that the *Journal of Experimental Psychology: General* declined to publish this paper. Therefore, these numerous, fundamental flaws continue to be in print in a top psychology journal. Papers in print are assumed to accurate unless stated otherwise. It is therefore disappointing to us that the journal did not remedy any subset of the numerous, fundamental problems that we describe above. We hope that our efforts will lead to more forthcoming behavior from journals in admitting and correcting their flawed publications.

References

Allred, S., Crawford, L.E., Duffy, S., & Smith, J. (2016). Working memory and spatial judgments: Cognitive load increases the central tendency bias. *Psychonomic Bulletin & Review, 23*(6), 1825-1831.

Ashourian, P., & Loewenstein, Y. (2011). Bayesian inference underlies the contraction bias in delayed comparison tasks. *PloS ONE, 6*(5), e19551.

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*(4), 744-763.

Barth, H., Lesser, E., Taggart, J., & Slusser, E. (2015). Spatial estimation: A non-Bayesian alternative. *Developmental Science, 18*(5), 853-862.

Blackwell, D., & Dubins, L. (1962). Merging of opinions with increasing information. *Annals of Mathematical Statistics, 33*, 882-886.

Bowers, J. S., & Davis, C. J. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*(3), 389-414.

Bowers, J. S., & Davis, C. J. (2012b). Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin, 138*(3), 423-426.

Cassey, P., Hawkins, G. E., Donkin, C., & Brown, S. D. (2016). Using alien coins to test whether simple inference is Bayesian. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(3), 497-503.

Chater, N., Goodman, N., Griffiths, T. L., Kemp, C., Oaksford, M., & Tenenbaum, J. B. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. *Behavioral and Brain Sciences, 34*(4), 194-196.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10*(7), 287-291.

Choplin, J. M., & Hummel, J. E. (2002). Magnitude comparisons distort mental representations of magnitude. *Journal of Experimental Psychology: General, 131*(2), 270-286.

Corbin, J. C., Crawford, L. E., & Vavra, D. T. (2017). Misremembering emotion: Inductive category effects for complex emotional stimuli. *Memory & Cognition, 45*(5), 691-698.

Corneille, O., Huart, J., Becquart, E., & Brédart, S. (2004). When memory shifts toward more typical category exemplars: Accentuation effects in the recollection of ethnically ambiguous faces. *Journal of Personality and Social Psychology, 86*(2), 236-250.

Crawford, L. E. (2019). Reply to Duffy and Smith's (2018) reexamination. *Psychonomic Bulletin & Review, 26*(2), 693-698.

Crawford, L. E., & Duffy, S. (2010). Sequence effects in estimating spatial location. *Psychonomic Bulletin & Review, 17*(5), 725-730.

Crawford, L. E., Huttenlocher, J., & Engebretson, P. H. (2000). Category effects on estimates of stimuli: Perception or reconstruction? *Psychological Science, 11*(4), 280-284.

Crosetto, P., Filippin, A., Katuščák, P., & Smith, J. (2019). When is a uniform not a uniform? The Central Tendency Bias in probability elicitations. Working paper, Rutgers University-Camden.

DeCarlo, L. T., & Cross, D. V. (1990). Sequential effects in magnitude scaling: Models and theory. *Journal of Experimental Psychology: General, 119*(4), 375-396.

Duffy, S., Huttenlocher, J., Hedges, L. V., & Crawford, L. E. (2010). Category effects on stimulus estimation: Shifting and skewed frequency distributions. *Psychonomic Bulletin & Review, 17*, 224-230.

Duffy, S., & Smith, J. (2018). Category effects on stimulus estimation: Shifting and skewed frequency distributions-A reexamination. *Psychonomic Bulletin & Review, 25*(5), 1740-1750.

Duffy, S., & Smith, J. (2019). Omitted-variable bias and other matters in the defense of the category adjustment model: A reply to Crawford. Working paper, Rutgers University-Camden.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193-242.

Elqayam, S., & Evans, J. S. B. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences, 34*(5), 233-248.

Eriksson, K. (2012). The nonsense math effect. *Judgment and Decision Making, 7*(6), 746-749.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*(2), 134-140.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116*(4), 752-782.

Fugate, J. M. (2013). Categorical perception for emotional faces. *Emotion Review,* 5(1), 84-89.

Goodman, N. D., Frank, M. C., Griffiths, T. L., Tenenbaum, J. B., Battaglia, P. W., & Hamrick, J. B. (2015). Relevant and robust: A response to Marcus and Davis (2013). *Psychological Science, 26*(4), 539-541.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin, 138*(3), 415-422.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17*(9), 767-773.

Hahn, U. (2014). The Bayesian boom: Good thing or bad? *Frontiers in Psychology, 5*, 765.

Hayes, K. J. (1953). The backward curve: A method for the study of learning. *Psychological Review, 60*(4), 269-275.

Hemmer, P., & Steyvers, M. (2009a). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin & Review, 16*(1), 80-87.

Hemmer, P., & Steyvers, M. (2009b). A Bayesian account of reconstructive memory. *Topics in Cognitive Science, 1*, 189-202.

Hemmer, P., Tauber, S., & Steyvers, M. (2015). Moving beyond qualitative evaluations of Bayesian models of cognition. *Psychonomic Bulletin & Review, 22*(3), 614-628.

Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2005). Judgments of risk frequencies: Tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(4), 621-642.

Holden, M. P., Curby, K. M., Newcombe, N. S., & Shipley, T. F. (2010). A category adjustment approach to memory for spatial location in natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 590-604.

Hollingworth, H. L. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods, 7*(17), 461-469.

Hund, A. M., & Spencer, J. P. (2003). Developmental changes in the relative weighting of geometric and experience-dependent location cues. *Journal of Cognition and Development, 4*(1), 3-38.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129*, 220-241.

Jesteadt, W., Luce, R. D., & Green, D. M. (1977). Sequential effects in judgments of loudness. *Journal of Experimental Psychology: Human Perception and Performance, 3*(1), 92-104.

Jones, M., Curran, T., Mozer, M. C., & Wilder, M. H. (2013). Sequential effects in response time reveal learning mechanisms and event representations. *Psychological Review, 120*(3), 628-666.

Jones, M., & Love, B. C. (2011a). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34*(4), 169-188.

Jones, M., & Love, B. C. (2011b). Pinning down the theoretical commitments of Bayesian cognitive models. *Behavioral and Brain Sciences, 34*(4), 215-231.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In Gilovich, T., Griffin, D., & Kahneman, D. (Eds.), *Heuristics and biases: The psychology of intuitive judgment*, Cambridge University Press, 49-81.

Kahneman, D., & Tversky, A., (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237-251.

Laming, D. (1984). The relativity of ‘absolute’ judgements. *British Journal of Mathematical and Statistical Psychology, 37*(2), 152-183.

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science, 33*(6), 969-998.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science, 24*(12), 2351-2360.

Marcus, G. F., & Davis, E. (2015). Still searching for principles: A response to Goodman et al. (2015). *Psychological Science, 26*(4), 542-544.

McCullough, S., & Emmorey, K. (2009). Categorical perception of affective and linguistic facial expressions. *Cognition, 110*(2), 208-221.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502–517.

Mozer, M. C., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science, 32*(7), 1133-1147.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357-395.

Olkkonen, M., & Allred, S. R. (2014). Short-term memory affects color perception in context. *PloS ONE, 9*(1), e86488.

Olkkonen, M., McCarthy, P. F., & Allred, S. R. (2014). The central tendency bias in color perception: Effects of internal and external noise. *Journal of Vision, 14*(11), 1-15.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition, 120*(3), 302-321.

Persaud, K., & Hemmer, P. (2014). The influence of knowledge and expectations for color on episodic memory. *Proceedings of the Cognitive Science Society, 36*, 1162-1167.

Petzold, P. (1981). Distance effects on sequential dependencies in categorical judgments. *Journal of Experimental Psychology: Human Perception and Performance, 7*(6), 1371-1385.

Petzold, P., & Haubensak, G. (2004). The influence of category membership of stimuli on sequential effects in magnitude judgment. *Perception & Psychophysics, 66*(4), 665-678.

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences, 19*(5), 285-293.

Poulton, E. C. (1979). Models for biases in judging sensory magnitude. *Psychological Bulletin, 86*(4), 777-803.

Psychology Software Tools, Inc. [E-Prime 2.0]. (2012). Retrieved from http://www.pstnet.com.

Rahnev, D. & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences, 41*, 1-66.

Roberson, D., Damjanovic, L., & Pilling, M. (2007). Categorical perception of facial expressions: Evidence for a "category adjustment" model. *Memory & Cognition, 35*(7), 1814-1829.

Sailor, K. M., & Antoine, M. (2005). Is memory for stimulus magnitude Bayesian? *Memory & Cognition, 33*, 840-851.

Sampson, R. J., & Raudenbush, S. W. (2004). Seeing disorder: Neighborhood stigma and the social construction of "broken windows." *Social Psychology Quarterly, 67*(4), 319-342.

Savage, L.J. (1954). *The Foundations of Statistics*. Wiley, New York. Reprinted in 1972 by Dover, New York.

Schutte, A. R., & Spencer, J. P. (2009). Tests of the dynamic field theory and the spatial precision hypothesis: Capturing a qualitative developmental transition in spatial working

memory. *Journal of Experimental Psychology: Human Perception and Performance, 35*(6), 1698-1725.

Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin, 49*(3), 263-269.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116*(3), 250-264.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Spencer, J. P., & Hund, A. M. (2002). Prototypes and particulars: Geometric and experience-dependent spatial categories. *Journal of Experimental Psychology: General, 131*(1), 16-37.

Spencer, J. P., & Hund, A. M. (2003). Developmental continuity in the processes that underlie spatial recall. *Cognitive Psychology, 47*(4), 432-480.

Sperber, D. (2010). The guru effect. *Review of Philosophy and Psychology, 1*(4), 583-592.

Staddon, J. E., King, M., & Lockhead, G. R. (1980). On sequential effects in absolute judgment experiments. *Journal of Experimental Psychology: Human Perception and Performance, 6*(2), 290-301.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702-712.

Stevens, S. S., & Greenbaum, H. B. (1966). Regression effect in psychophysical judgment. *Perception & Psychophysics, 1*(5), 439-446.

Stewart, N., Brown, G. D., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 3-11.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review, 124*(4), 410-441.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*(7), 309-318.

Twedt, E., Crawford, L. E., & Proffitt, D. R. (2015). Judgments of others' heights are biased toward the height of the perceiver. *Psychonomic Bulletin & Review, 22*(2), 566-571.

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J. S., Renaut, S., & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology, 24*(1), 94-97.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology, 100*(3), 426-432.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832.

Wilder, M., Jones, M., & Mozer, M. C. (2009). Sequential effects reflect parallel learning of multiple environmental regularities. *Advances in Neural Information Processing Systems, 22*, 2053-2061.

Woods, A. T., Poliakoff, E., Lloyd, D. M., Dijksterhuis, G. B., & Thomas, A. (2010). Flavor expectation: The effect of assuming homogeneity on drink perception. *Chemosensory Perception, 3*(3-4), 174-181.

Young, S. G., Hugenberg, K., Bernstein, M. J., & Sacco, D. F. (2009). Interracial contexts debilitate same-race face recognition. *Journal of Experimental Social Psychology, 45*(5), 1123-1126.

Yu, A. J., & Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? *Advances in Neural Information Processing Systems, 21*, 1873-1880.

Supplemental Online Appendix

A1. Running mean regressions, fixed-effects analysis

The analysis summarized in Table 3 finds only weak evidence that the running mean is related to response. However, the reader might be concerned that the results are not robust to the specification of the repeated nature of the data. Here we perform an analysis with the same independent variables but we offer a different repeated measures specification. We do not assume a correlation between judgments by the same participant, but rather we account for the heterogeneity by estimating a unique intercept for each participant. In other words, rather than running random-effects regressions, here we run fixed-effects regressions. These regressions are summarized in Table A1.

Table A1: Fixed-effects repeated measures regressions of the response variable

	Normal	Uniform	Short	Long
Target	0.765***	0.753***	0.833***	0.730***
	(0.008)	(0.008)	(0.008)	(0.014)
Running mean	0.147*	0.082	-0.008	0.063
	(0.069)	(0.059)	(0.080)	(0.122)
-2 Log L	18221.7	16639.9	18684.8	20341.1
Observations	1882	1680	2095	2056

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts or the participant dummies. † indicates significance at $p < .1$, * indicates significance at $p < .05$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

Similar to Table 3, here target is significantly related to response in every specification. Also similar to Table 3, running mean is only significant in the normal treatment specification.

A2. Preceding targets, fixed-effects analysis

Table 4 reports that, in every treatment, previous target is significantly related to response whereas running mean is not related to response. This analysis was conducted with a random-effects analysis. Here we perform the analysis with fixed-effects regressions. These regressions are summarized in Table A2.

Table A2: Fixed-effects repeated measures regressions of the response variable

	Normal	Uniform	Short	Long
Target	0.766***	0.753***	0.835***	0.735***
	(0.008)	(0.008)	(0.008)	(0.014)
Running mean	0.102	0.049	-0.095	-0.019
	(0.070)	(0.060)	(0.080)	(0.123)
Previous target	0.030***	0.025***	0.053***	0.058***
	(0.008)	(0.008)	(0.008)	(0.014)
-2 Log L	18215.9	16637.6	18653.4	20330.9
Observations	1882	1680	2095	2056

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts or the participant dummies. † indicates significance at $p < .1$ and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

We note that the results in Table A2 are nearly identical to that in Table 4.

The analysis summarized in Table 5 finds that the preceding targets variable offers a better prediction of response variable than running mean. In other words, rather than running random-effects regressions, here we run fixed-effects regressions. Table A3 summarizes this fixed-effects analysis.

Table A3: Fixed-effects repeated measures regressions of the response variable.

	None	Prec 1	Prec 3	Prec 5	Prec 10
Target	0.765***	0.766**	0.766***	0.766***	0.766***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Running mean	0.0869*	0.0381	0.0272	0.0436	0.0291
	(0.0368)	(0.0372)	(0.0384)	(0.0398)	(0.0436)
Preceding targets	-	0.0343***	0.0445***	0.0331**	0.0479*
		(0.0045)	(0.0084)	(0.0117)	(0.0193)
-2 Log L	74477.8	74428.4	74457.4	74476.8	74477.7

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts, the treatment dummies, or the participant dummies. All regressions have 7713 observations. † indicates significance at $p < .1$, * indicates significance at $p < .05$, ** indicates significance at $p < .01$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

Just as in Table 5, preceding targets is significant in every specification, and running mean is not significant in any specification that accounts for the previous lines.

A3. Simulated response35 variable

In Table 6 we analyzed the simulated Respone25 variable. Here we perform the identical analysis with the simulated response35 variable, which contains noise with a standard deviation of 35 pixels, rather than 25 pixels. Table A4 summarizes this analysis.

Table A4: Random-effects repeated measures regressions of the simulated response35 variable.

	No Prec	Prec 1	Prec 3	Prec 5	Prec 10
Target	0.921***	0.921***	0.921***	0.921***	0.921***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Running mean	0.112**	0.111**	0.113**	0.130**	0.118**
	(0.037)	(0.038)	(0.039)	(0.040)	(0.043)
Preceding targets	-	0.0012	-0.0008	-0.0144	-0.0052
		(0.0047)	(0.0087)	(0.0121)	(0.0199)
-2 Log L	76560.9	76569.7	76568.6	76566.5	76566.8

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts, the treatment dummies or the covariance parameters. All regressions have 7831 observations. † indicates significance at $p < .1$, ** indicates significance at $p < .01$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

In every specification, running mean is significant at .01 and in none of the specifications is preceding targets significant. We also note that a fixed-effects analysis, rather than a random-effects analysis, does not change these results.

We note that the noise in the analysis of Table A4 exceeds that in our original analysis in Table 4, as can be seen by comparing the -2 Log L values. We also note that the noise in the analysis of Table 6 is less than that in the analysis of Table 5, as can be seen by comparing the -2 Log L values. Given the results of Tables 6 and A4, we reject the criticism that the declining standard deviation of running mean prevents satisfactory estimates of the coefficient of the running mean variable. Further, whereas Table 6 and Table A4 perform a random-effects analysis, we also perform fixed-effects versions of these analyses and the results are not changed.

A4. Responses with zero mass across trials

We conduct an analysis similar to Table 7, but Table A5 summarizes this on only the first half of trials. There are 250 responses with a zero mass and 3639 without. As it would not be identified, we do not include the First half specification.

Table A5: Fixed-effects logistic regressions of the zero mass dummy variable.

	Trial	Sqrt. Trial	First 5	First 10	First 20
Trial	0.0005	0.0025	-0.072	0.063	0.178
	(0.0028)	(0.0339)	(0.365)	(0.258)	(0.183)
-2 Log L	1163.5	1163.5	1163.5	1163.4	1162.6

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 1 through 96. We do not provide the estimates of the intercepts, the participant dummy variables, or the treatment-target dummy variables. All regressions have 3889 observations. † indicates significance at $p < .1$. -2 Log L refers to negative two times the log-likelihood.

Similar to the results in Table 7, here we do not find evidence that zero mass responses became less likely across trials. This suggests that participants either did not learn this aspect of the distribution or they did not use this to inform their judgments.

A5. Bias toward the running mean across trials

Table A6 was performed as Table 8, with the running mean bias as the dependent variable, but on only the first half of trials.

Table A6: Random-effects regressions of the running mean bias variable.

	Trial	Sqrt. Trial	First 5	First 10	First 20
Trial	0.061***	0.803***	-6.404**	-4.150**	-4.065***
	(0.016)	(0.196)	(2.205)	(1.501)	(1.089)
-2 Log L	35919.9	35913.2	35916.6	35918.1	35912.5

Notes: We provide the coefficient estimates with the standard errors in parentheses. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. We examine trials 2 through 96. All regressions have 3851 observations. ** indicates significance at $p < .01$ and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

Just as in Table 8, here we find strong evidence that the bias toward the running mean increases across trials.

The reader is possibly concerned that the running mean bias variable is not sufficiently close to the weight between the running mean and the noisy memory (λ). Therefore, we define the running mean bias ratio to be the distance between the target and the running mean divided by the sum of the distance between the target and the running mean and the distance between the response and the running mean:

$$\text{Running mean bias ratio} = |\text{Target} - \text{Running mean}| / [\,|\text{Target} - \text{Running mean}| + |\text{Response} - \text{Running mean}|\,].$$

Table A7 was performed as Table 8, but with the running mean bias ratio on all trials.[51]

Table A7: Random-effects regressions of the running mean bias ratio variable.

	Trial	Sqrt. Trial	First 5	First 10	First 20
Trial	0.00014***	0.00267**	-0.033**	-0.026**	-0.026***
	(0.00003)	(0.00056)	(0.013)	(0.009)	(0.006)
-2 Log L	6214.5	6224.9	6215.3	6216.7	6225.5

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 7712 observations. ** indicates significance at $p < .01$ and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

Similar to that in Table 8, here we find strong evidence that the bias toward the running mean increases across trials.

Table A8 was performed as Table 9, with previous bias as independent variable, but on only the first half of trials.

Table A8: Random-effects regressions of the previous bias variable.

	Trial	Sqrt. Trial	First 5	First 10	First 20
Trial	0.064***	0.875***	-9.253***	-6.610***	-4.389***
	(0.019)	(0.238)	(2.669)	(1.817)	(1.319)
-2 Log L	37353.8	37346.2	37342.8	37342.4	37345.2

Notes: We provide the coefficient estimates with the standard errors in parentheses. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy

[51] One observation was such that the running mean was equal to both the target and the response, thus implying an undefined running mean bias ratio. Therefore we have one fewer observation in Table A7 than in Table 7.

variables. We examine trials 2 through 96. All regressions have 3851 observations. *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

As with Table 9, we see that there is an increase in the bias toward the previous line across trials. We also note that a fixed-effects specification does not change these results.

A6. Error across trials

We now perform an analysis, identical to that in Table 10, but restricted to the first half of trials. Table A9 summarizes this analysis. We note that CAM would predict a negative estimate for Trial and positive estimates for the others.

Table A9: Random-effects logistic regressions of the absolute response bias variable.

	Trial	Sqrt. Trial	First 5	First 10	First 20
Trial	0.070***	0.797***	-0.569	-2.467*	-3.494***
	(0.013)	(0.158)	(1.653)	(1.192)	(0.886)
-2 Log L	34881.9	34881.0	34901.6	34898.1	34887.4

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 1 through 96. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 3889 observations. † indicates significance at $p < .1$, * indicates significance at $p < .05$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

Similar to Table 10, here we find that absolute response bias is increasing in the Trial, First 10, and First 20 specifications.

A7. On the non-linear relationship of the data

HHV write on pages 228-229, "In order to determine whether the shape of the bias curve in the normal and uniform conditions is different, it is desirable to compare a numerical index of bias shape….One such index of linearity is obtained by considering the problem as a special case of repeated measures and using orthogonal polynomial contrasts. For the uniform and normal treatments, we performed an analysis of variance using each participant's mean response to each stimulus value and treating the stimuli as different levels of a factor repeated within participants.

A sum of squares was calculated for each orthogonal polynomial component; the degree to which the linear component accounts for total variability was estimated by η^2, the ratio of the linear component's sum of squares to the total of the linear and nonlinear components' sums of squares. Values of η^2 approaching 1.0 indicate that the bias pattern is primarily linear, whereas lower values indicate departures from linearity."

This measure of non-linearity seems similar to an R^2 but averaged across observations within participants and targets. Further, it would seem that this measure would not be able to distinguish between a non-linear polynomial and simply a noisy linear relationship. Finally, we note that the prediction given in Figure 2B is that the relationship will have a shape similar to a quadratic.

Therefore, we include a term that accounts for this possible quadratic relationship. We perform the analysis as in Table 3, but we also include a variable Target squared variable. These regressions are summarized in Table A10.

Table A10: Random-effects repeated measures regressions of the response variable

	Normal	Uniform	Short	Long
Target	0.796***	0.872***	0.939***	0.541**
	(0.040)	(0.036)	(0.047)	(0.185)
Running mean	0.149*	0.085	-0.012	0.061
	(0.068)	(0.059)	(0.080)	(0.122)
Target squared	-0.00007	-0.0003***	-0.0004*	0.0003
	0.00008	0.00008	(0.0002)	(0.0003)
-2 Log L	18306.4	16709.6	18757.0	20446.7
Observations	1882	1680	2095	2056

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts or the covariance parameters. † indicates significance at $p < .1$, * indicates significance at $p < .05$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

The inclusion of the Target squared variable does not affect the conclusion from Table 3 that there is only weak evidence of a bias toward the running mean.

We perform the analysis as summarized in Table 4, but we include the Target Squared variable. These regressions are summarized in Table A11.

Table A11: Random-effects repeated measures regressions of the response variable

	Normal	Uniform	Short	Long
Target	0.795***	0.872***	0.932***	0.558**
	(0.039)	(0.036)	(0.047)	(0.185)
Running mean	0.104	0.052	-0.098	-0.020
	(0.069)	(0.060)	(0.080)	(0.123)
Previous target	0.030***	0.025**	0.052***	0.057***
	(0.008)	(0.008)	(0.008)	(0.014)
Target squared	-0.00006	-0.0003***	-0.0004*	0.0003
	(0.00008)	(0.00008)	(0.0002)	(0.0003)
-2 Log L	18300.7	16707.3	18726.4	20436.6
Observations	1882	1680	2095	2056

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts or the covariance parameters. † indicates significance at $p < .1$, ** indicates significance at $p < .01$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

The inclusion of the Target squared variable does not change the conclusions from Table 4, that there is a bias toward previous targets but not a bias toward the running mean.

We perform the analysis as in Table 5, but we include the Target squared variable. These regressions are summarized in Table A12.

Table A12: Random-effects repeated measures regressions of the response variable

	None	Prec 1	Prec 3	Prec 5	Prec 10
Target	0.868***	0.868***	0.870***	0.870***	0.870***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
Running mean	0.089*	0.041	0.029	0.044	0.030
	(0.037)	(0.037)	(0.038)	(0.040)	(0.043)
Preceding targets	-	0.034***	0.045***	0.034**	0.049*
		(0.004)	(0.008)	(0.012)	(0.019)
Target squared	-0.0002***	-0.0002***	-0.0002***	-0.0002***	-0.0002***
	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.00003)
-2 Log L	74758.1	74709.2	74737.5	74756.5	74757.7

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts, the treatment dummies, or the covariance parameters. All regressions have 7713 observations. † indicates significance at $p <$

.1, * indicates significance at $p < .05$, ** indicates significance at $p < .01$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

Again, the inclusion of the Target squared variable does not alter the conclusions of Table 5 that there is a bias toward recent targets but not toward the running mean.

The reader might worry that these results are an artifact that in the analysis summarized in Table A12 we restricted the relationship involving Target and Response to be identical across all four treatments. We perform the analysis, as in Table A12, but we allow the relationship between target and response to vary by treatment. The Target estimate corresponds to the Long treatment. The interactions with the treatments and target correspond to the differences between that in the treatment and in the Long treatment. These regressions are summarized in Table A13.

Table A13: Random-effects repeated measures regressions of the response variable

	None	Prec 1	Prec 3	Prec 5	Prec 10
Target	0.841***	0.842***	0.845***	0.845***	0.844***
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
Target*Normal	0.002	0.0006	0.00001	-0.0006	0.0005
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Target*Uniform	-0.010	-0.012	-0.012	-0.012	-0.011
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
Target*Short	0.039	0.038	0.038	0.037	0.038
	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)
Running mean	0.091*	0.042	0.030	0.044	0.030
	(0.037)	(0.037)	(0.038)	(0.040)	(0.043)
Preceding targets	-	0.034***	0.045***	0.036**	0.050**
		(0.004)	(0.008)	(0.012)	(0.019)
Target squared	-0.0002***	-0.0002**	-0.0002***	-0.0002***	-0.0002***
	(0.00005)	(0.00005)	(0.00005)	(0.00005)	(0.00005)
-2 Log L	74768.7	74718.7	74747.5	74766.4	74768.0

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts, the treatment dummies, or the covariance parameters. All regressions have 7713 observations. † indicates significance at $p < .1$, * indicates significance at $p < .05$, ** indicates significance at $p < .01$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

Including a unique estimate for target for every treatment does not alter our conclusion that there is a bias toward the recent targets but not toward the running mean.

The reader might worry that these results are an artifact that the analysis summarized in Table A13 restricted the relationship involving Target squared and Response to be identical across all four treatments. We perform the analysis, as in Table A13, but we allow the relationship between Target squared and response to vary by treatment. The Target squared estimate corresponds to the Long treatment. The interactions with the treatments and Target squared correspond to the differences between that in the treatment and in the Long treatment. These regressions are summarized in Table A14.

Table A14: Random-effects repeated measures regressions of the response variable

	None	Prec 1	Prec 3	Prec 5	Prec 10
Target	0.540**	0.551***	0.546***	0.543***	0.543***
	(0.164)	(0.163)	(0.164)	(0.164)	(0.164)
Target * Normal	0.254	0.243	0.252	0.254	0.253
	(0.169)	(0.168)	(0.168)	(0.168)	(0.168)
Target * Uniform	0.331*	0.321†	0.327*	0.331*	0.331*
	(0.167)	(0.166)	(0.167)	(0.167)	(0.167)
Target * Short	0.400*	0.384*	0.394*	0.399*	0.399*
	(0.178)	(0.177)	(0.177)	(0.177)	(0.178)
Running mean	0.091*	0.042	0.031	0.044	0.030
	(0.037)	(0.037)	(0.038)	(0.040)	(0.043)
Preceding targets	-	0.034***	0.045***	0.036**	0.050**
		(0.004)	(0.008)	(0.012)	(0.019)
Target squared	0.0003	0.0003	0.0003	0.0003	0.0003
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Target squared *	-0.0004	-0.0003	-0.0004	-0.0004	-0.0004
Normal	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Target squared *	-0.0006*	-0.0005*	-0.0005*	-0.0006*	-0.0006*
Uniform	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Target squared *	-0.0007†	-0.0007†	-0.0007†	-0.0007†	-0.0007†
Short	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
-2 Log L	74807.1	74757.4	74786.2	74804.9	74806.3

Notes: We provide the coefficient estimates with the standard errors in parentheses. We examine trials 2 through 192. We do not provide the estimates of the intercepts, the treatment dummies, or the covariance parameters. All regressions have 7713 observations. † indicates significance at $p < .1$, * indicates significance at $p < .05$, ** indicates significance at $p < .01$, and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

Again, not surprisingly, including a unique estimate for target squared for every treatment does not alter our conclusion that there is a bias toward the recent targets but not toward the running mean.