



Munich Personal RePEc Archive

**Category effects on stimulus estimation:
Shifting and skewed frequency
distributions - A reexamination**

Duffy, Sean and Smith, John

Rutgers University-Camden

6 September 2017

Online at <https://mpra.ub.uni-muenchen.de/81254/>
MPRA Paper No. 81254, posted 09 Sep 2017 04:53 UTC

Category effects on stimulus estimation: Shifting and skewed frequency distributions
-A reexamination*

Sean Duffy

Rutgers University-Camden

John Smith

Rutgers University-Camden

Address correspondence to:

John Smith

Department of Economics
Rutgers University-Camden

311 N. 5th Street

Camden, NJ

08102 USA

smithj@camden.rutgers.edu

September 5, 2017

*We thank Roberto Barbera, I-Ming Chiu, L. Elizabeth Crawford, Johanna Hertel, Rosemarie Nagel, and Adam Sanjurjo for helpful comments. This project was supported by Rutgers University Research Council Grant #202297. John Smith thanks Biblioteca de Catalunya.

Abstract

Duffy, Huttenlocher, Hedges, and Crawford (2010) [*Psychonomic Bulletin & Review*, 17(2), 224-230] report on experiments where participants estimate the lengths of lines. These studies were designed to test the Category Adjustment Model (CAM), a Bayesian model of judgments. The authors report that their analysis provides evidence consistent with CAM: that there is a bias toward the running mean and not recent stimuli. We reexamine their data. First, we attempt to replicate their analysis and we obtain different results. Second, we conduct a different statistical analysis. We find significant recency effects and we identify several specifications where the running mean is not significantly related to judgment. Third, we conduct tests of auxiliary predictions of CAM. We do not find evidence that the bias towards the mean increases with exposure to the distribution. We also do not find that responses longer than the maximum of the distribution or shorter than the minimum become less likely with greater exposure to the distribution. Fourth, we produce a simulated dataset that is consistent with key features of CAM and our methods correctly identify it as consistent with CAM. We conclude that the Duffy et al. (2010) dataset is not consistent with CAM. We also discuss how conventions in psychology do not sufficiently reduce the likelihood of these mistakes in future research. We hope that the methods that we employ will be used to evaluate other datasets.

Keywords: judgment, memory, Category Adjustment Model, central tendency bias, recency effects, Bayesian judgments

A well-known experimental effect is that participants tend to make judgments biased toward the mean of the distribution of stimuli. This experimental effect is often referred to as the *central tendency bias* (Hollingworth, 1910; Goldstone, 1994).

The Category Adjustment Model, hereafter referred to as CAM, offers a Bayesian explanation for this effect. CAM (Huttenlocher, Hedges, & Vevea, 2000)¹ holds that participants imperfectly perceive and remember stimuli. According to CAM, in order to compensate for these imperfections, participants improve accuracy by considering information about the probability distribution of the stimuli according to Bayes' rule. In particular, CAM suggests that judgments will be a weighted average of the imperfect memory of the stimulus and the mean of the distribution of previously seen stimuli. The weighted average is a function of the standard deviation of the distribution and the standard deviation of the noisy memory. These weights are optimal in that the resulting judgments minimize the error, however they also produce judgments consistent with the central tendency bias.

Since judgments consistent with CAM minimize the errors associated with limited memory and imperfect perception, CAM predicts that judgments will not be affected by features of the experiment that do not improve the accuracy of the judgment. In particular, one prediction of CAM is that participants will not be sensitive to recently viewed stimuli. Another prediction of CAM is that there will be a negative relationship between the central tendency bias of judgments and the standard deviation of the distribution of stimuli.

Huttenlocher et al. (2000) conducted three experiments to test CAM. Participants performed serial judgment tasks on the fatness of computer generated images of fish, on the shades of grey, and on the lengths of lines. In each of these settings, participants performed judgments under four distributions of stimuli, which exhibit different means and standard

¹ See Xu and Griffiths (2010) for a similar model.

deviations. Huttenlocher et al. (2000) analyze all three experiments in a similar manner: they examine data averaged across trials. The authors conclude that CAM is “verified” and that other explanations, such as a bias toward only a set of recent stimuli, cannot explain the data.

CAM has had a large impact on the judgment literature and has influenced research in topics as disparate 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), judgments of likelihood (Hertwig, Pachur, & Kurzenhäuser, 2005), and facial recognition (Corneille, Huart, Becquart, & Brédart, 2004).

Duffy, Huttenlocher, Hedges, and Crawford (2010), hereafter referred to as DHHC, study whether judgments of the lengths of lines are consistent with CAM for distributions that are asymmetric or shifting. Similar to Huttenlocher et al. (2000), DHHC largely analyze averaged data. In their abstract, DHHC state, “...we find that people adjust estimates toward the category’s running mean, which is consistent with the CAM but not with alternative explanations for the adjustment of stimuli toward a category’s central value.” On page 229, DHHC state that their results “...provide direct evidence that the central tendency bias in stimulus estimation cannot be explained as a memory blend between the magnitude of a target stimulus and a small set of stimuli immediately preceding it.” These conclusions are the focus of this reexamination.

Although we will say more about this below, at the outset it should be noted that the first author on DHHC is also the first author on this reexamination. Our reexamination is the product of the first author, who had the original data, and the second author, who expressed skepticism toward the DHHC results. Also note that we subsequently refer to the authors of DHHC in third

person, despite that the sets of coauthors on DHHC and this reexamination are not disjoint. We will say more on this matter as it becomes relevant.

DHHC Experiment 1

DHHC state, “The purpose of Experiment 1 was to determine whether participants adjusted responses toward the mean of all stimuli presented or to some other point such as the mean of a small number of recent stimuli from a recent subset (e.g., the last 1, 2, 3, ..., 10 stimuli).”

Description of Methods

Participants were directed to judge the length of lines with 19 possible stimulus sizes, ranging from 80 to 368 pixels, in increments of 16 pixels. We refer to the line that is to be estimated as the *target line*.

Participants were presented with the target line then the target line disappeared. Subsequently an initial adjustable line appeared. The participant would manipulate the length of this adjustable line until they judged its length to be that of the target line. We refer to this response as the *response line*. DHHC report that roughly half of the participants had an initial adjustable line of 40 pixels and the other half had an initial adjustable line of 400 pixels.

Participants estimated target lines from one of two distributions. Consider labeling the targets 1 through 19, such that they are increasing length. In the Right skew distribution (long lines less likely than short lines)² participants were shown 9 instances of targets 1 and 2, 8 instances of targets 3 and 4, and so on, to 5 instances of targets 9, 10, and 11, 4 instances of targets 12 and 13, and so on, to 1 instance of targets 18 and 19. These lines were drawn at random without replacement. In the Left skew distribution (long lines more likely than short lines) participants were shown 9 instances of targets 18 and 19, 8 instances of targets 16 and 17,

² We follow the counter-intuitive convention that the direction refers to the tail and not the mode

and so on, to 5 instances of targets 9, 10, and 11, 4 instances of targets 7 and 8, and so on, to 1 instance of targets 1 and 2. Again, these lines were drawn at random without replacement.

In one treatment, participants estimated the length of the 95 lines drawn from the Left skew distribution then, without announcement, estimated the 95 lines drawn from the Right skew distribution. In the other treatment, participants estimated the length of the 95 lines drawn from the Right skew distribution then, without announcement, estimated the 95 lines drawn from the Left distribution. Each participant therefore was exposed to the identical set of 190 lines.

The study had 25 participants therefore the total number of judgments was 4750. The reader is referred to DHHC for further details.³

Description of the dataset

Among these 4750 observations, there are 30 missing values for the response line. It seems as if the DHHC authors removed these observations because the responses were below the lower bound of possible responses. We note that these 30 observations account for less than 1% of the total observations and we expect that they would not affect the analysis.

We also note that the dataset does not possess information about the initial adjustable line length. This is regrettable because Allred et al. (2016) find evidence that the initial adjustable line affects judgments of length. We are therefore not able to determine the effect of these initial adjustable line lengths on the response.

Additionally, we note that the randomization in the experiment was not completely satisfactory. We find a negative correlation between the target line and the trial number in the first 95 trials of the Right skew then Left skew treatment ($r(1140) = -.076, p = .01$). To our knowledge no other such correlation exists. Although we note that CAM would predict that such a serial correlation would not affect judgments.

³ We include this dataset and the code from our analysis in the Supplemental Online Appendix.

Analysis in DHHC

DHHC report that they performed the following regressions on each participant with Response as the dependent variable. The independent variables were the target line length, the running mean of the previous target lines, and the average of the preceding 20 target lines.

DHHC estimate the coefficients (β) for the following specification:

$$\text{Response} = \beta_1(\text{Target}) + \beta_2(\text{Running mean}) + \beta_3(\text{Preceding 20 targets}).$$

Regarding the Running mean variable, DHHC report that, “The mean of all stimuli (Running mean) has a shorter but statistically significant impact in all cases...” Regarding the Preceding 20 targets variable, the authors state, “The impact of the preceding 1 to 20 stimuli is much shorter and statically insignificant ($p > .1$) in every analysis.”

The authors conclude their discussion of Experiment 1, by claiming that their results are consistent with CAM. In fact, in regards to the results from Experiment 1, DHHC state on page 227, “They are inconsistent with accounts arguing that the central tendency bias is a distortion caused by the immediate preceding stumuli.”

Our reexamination

Before we begin our reexamination, we say a few words about DHHC. Despite the overlap in authorship, there are certain details of the data and the analysis that are not reported in DHHC and the first author of this reexamination cannot recall.

For instance, the dataset has 30 missing values for the Response variable. Since the authors did not report the number of observations in their regressions, it is not possible to determine if the analysis was conducted with these missing values.

Further, DHHC did not precisely specify how the Preceding 20 target variable was calculated. It is possible that observations without each of the 20 previous target lines (for

instance, the third judgment) were ignored. On the other hand, it is possible that this variable was calculated by considering as many available previous observations as possible, but constrained to not be more than 20. Since DHHC did not report the number of observations in the regressions, we cannot infer their method. In order to use all available data, we employ the latter of these methods.

It seems from the description of the analysis that DHHC estimated a specification⁴ in which the intercept was assumed to be zero. However, the authors did not justify this rather strong assumption. Regardless, we replicate their analysis by conducting the regressions both with the assumption of a zero intercept and with the assumption that the intercept is not constrained to be zero. We summarize these regressions in Table 1.

Table 1: Distribution of p-values of the Running mean and the Preceding 20 targets.

		Preceding 20 targets p-values					Total
		$p \geq .1$	$.1 > p \geq .05$	$.05 > p \geq .01$	$.01 > p \geq .001$	$.001 > p$	
Running mean p-values	$p \geq .1$	6 (15)	1 (2)	4 (5)	0 (1)	0 (0)	11 (23)
	$.1 > p \geq .05$	2 (0)	1 (1)	0 (0)	0 (0)	0 (0)	3 (1)
	$.05 > p \geq .01$	3 (0)	0 (0)	0 (0)	0 (0)	1 (0)	4 (0)
	$.01 > p \geq .001$	2 (0)	0 (0)	0 (0)	0 (0)	0 (1)	2 (1)
	$.001 > p$	3 (0)	1 (0)	1 (0)	0 (0)	0 (0)	5 (0)
	Total	16 (15)	3 (3)	5 (5)	0 (1)	1 (1)	25 (25)

Notes: We conduct a regression for each of the 25 participants. We report the p-values for the Running mean variable and the Preceding 20 variable. The specification with a zero intercept is reported outside the parentheses and the specification where the intercept is not constrained to be zero is reported inside the parentheses. Due to the incomplete data, each regression is conducted with observations that range from 182 to 189. Both specifications are conducted with a total of 4696 observations.

In contrast to the results reported by DHHC, our analysis suggests that the Preceding 20 targets variable is significant in 6 (24%) of the regressions with an assumed zero intercept and it is significant in 7 (28%) of the regressions where the intercept is not constrained to be zero.

Further, in contrast to the analysis of DHHC, we find that the Running mean variable is not

⁴ 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, and the assumptions regarding the error term.

significant in 14 (56%) of the regressions that assume a zero intercept and it is not significant in 24 (96%) of the regressions where the intercept is not constrained to be zero.⁵

We admit that an error in the execution of the analysis is likely responsible for the results presented by DHHC. Regrettably, Duffy (the first author on DHHC) cannot recall the origin of the erroneous regressions. Regardless of the origin of the mistaken analysis, our analysis suggests drastically different conclusions than those given in DHHC. In particular, we find that there are recency effects and that many participants do not exhibit a significant relationship between the Running mean and the Response. In summary, employing the technique used by DHHC, we do not find evidence in support of CAM.

Repeated measures regressions for preceding target lines

Above we attempted to replicate the findings of DHHC using their technique, however these methods would seem to not be ideal. For instance, their analysis does not provide an aggregate estimate of the relationships among the variables. Additionally, running participant-level regressions renders the summary of the analyses to be needlessly cumbersome. Finally, the balance of the analysis of DHHC analyzes averaged data, and this renders it difficult to distinguish between a bias toward the running mean and a bias toward recent stimuli.⁶

Here we employ standard repeated measures techniques in order to remedy these shortcomings. Since every response has an associated target, running mean, and set of recent targets, we include each of these variables in our analysis.

⁵ We note that the possible assumption of a zero intercept would be questionable given that no justification was provided and that with a non-zero intercept, only 1 participant has a significant relationship with the Running mean.

⁶ The drawbacks of analyzing averaged data have been known for some time (Estes, 1956; Siegler, 1987) and such concerns even appear in the recent judgments literature (Cassey et al., 2016; Hemmer, Tauber, & Steyvers, 2015).

We note that recency effects and sequential effects have been studied in the literature.⁷ As the analysis above, we include a specification that has an independent variable that is the average of the preceding 20 target lines. We refer to this variable as *Prec 20*. Since it is not obvious to us why the previous 20 targets were analyzed rather than other numbers, we run different specifications that include different numbers of preceding targets. We include a specification that accounts for only the preceding target line, which we refer to as *Prec 1*. Additionally, we calculate the average of the preceding 3, the preceding 5, the preceding 10, and the preceding 15 target lines. We refer to these variables, respectively, as *Prec 3*, *Prec 5*, *Prec 10*, and *Prec 15*. Our analysis below considers each of these 6 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.

Further, 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. Table 2 summarizes this random-effects analysis.⁸

⁷ See Choplin and Hummel (2002), DeCarlo and Cross (1990), Jones et al. (2013), Petzold and Haubensak (2004), Stewart, Brown, and Chater (2002), Wilder, Jones, and Mozer (2009), and Yu and Cohen (2009).

⁸ We note that Table 2 and the regression tables that follow are not consistent with the 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.

Table 2: Random-effects repeated measures regressions of the Response variable.

	No Prec	Prec 1	Prec 3	Prec 5	Prec 10	Prec 15	Prec 20
Target	0.813*** (0.005)	0.808*** (0.005)	0.803*** (0.005)	0.802*** (0.005)	0.804*** (0.005)	0.805*** (0.005)	0.807*** (0.005)
Running mean	0.200*** (0.024)	0.115*** (0.025)	0.040 (0.028)	0.037 (0.030)	0.046 (0.034)	0.056 (0.037)	0.065 (0.040)
Preceding targets	-	0.050*** (0.005)	0.093*** (0.009)	0.094*** (0.011)	0.088*** (0.014)	0.081*** (0.016)	0.075*** (0.018)
-2 Log L	45254.3	45179.3	45154.7	45187.6	45221.5	45235.3	45242.7

Notes: We provide the coefficient estimates with the standard errors in parentheses. We do not provide the estimates of the intercepts or the covariance parameters. All regressions have 4696 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, we find that the Preceding targets variable is significantly related to the response of the participant. Additionally, we see that the Running mean variable is significant only in the specifications without any preceding targets and with information about only the previous target. However, in the other 5 specifications, the Running mean variable is not significantly related to the Response.⁹ This suggests that the Preceding targets variables tend to be a better predictor of Response than Running mean, in contradiction with CAM and the stated results of DHHC.¹⁰ Further, we note that the DHHC draws were done without replacement. Thus, observing a particular target line implies that it is less likely to appear in the future. Therefore, finding recency effects stand in stark contrast to the predictions of CAM.

Some researchers might suspect that the above analysis is not sufficiently sensitive to detect evidence of CAM. In particular, a researcher might note that the standard deviation of the Running mean variable decreases across trials and this might prevent a satisfactory inference of the coefficient of the Running mean variable. In order to investigate this possibility, we simulated a simple dataset that is consistent with a key feature CAM and has parameters similar

⁹ In the Supplemental Online Appendix, we report Table A1, which summarizes the analogous analysis, but with fixed-effects, not random-effects. There are no qualitative differences between the results.

¹⁰ We employed heteroscedasticity robust standard errors (hccme=2 and hccme=4 in the panel procedure in SAS) in the analysis similar to that in Table 2 and our results are unchanged.

to that found in the DHHC data. We took the sequence of Target lines from Experiment 1 and added normally distributed noise, with a zero mean and a standard deviation of 25 pixels to each Target line. We refer to the sum of the 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 roughly 80% of the weight was placed on the memory of the target line, here we put 90% of the weight on Memory:

$$\text{Response25} = .9(\text{Memory}) + .1(\text{Running mean}).$$

These simulated judgments are clearly consistent with a key feature of CAM in that Response25 is biased toward Running mean but not toward recent lines. Additionally, there is a lower weight on the Running mean variable than in the dataset analyzed in Table 2. Therefore, detecting a relationship between Running mean and Response is more difficult in our simulated data than in the DHHC data. We perform the identical analysis to that performed in Table 2, which we summarize in Table 3.

Table 3: Random-effects repeated measures regressions of the simulated Response25 variable.

	No Prec	Prec 1	Prec 3	Prec 5	Prec 10	Prec 15	Prec 20
Target	0.899*** (0.004)	0.899*** (0.004)	0.898*** (0.004)	0.898*** (0.004)	0.897*** (0.004)	0.898*** (0.004)	0.899*** (0.004)
Running mean	0.105*** (0.007)	0.106*** (0.007)	0.105*** (0.007)	0.104*** (0.007)	0.103*** (0.008)	0.104*** (0.008)	0.105*** (0.008)
Preceding targets	-	-0.003 (0.004)	0.002 (0.006)	0.004 (0.007)	0.006 (0.008)	0.004 (0.008)	0.001 (0.009)
-2 Log L	42607.7	42616.3	42616.1	42615.5	42614.9	42615.2	42615.4

Notes: We provide the coefficient estimates with the standard errors in parentheses. We do not provide the estimates of the intercepts or the covariance parameters. All regressions have 4696 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, the Running mean variable is significant at .001 and the Preceding targets variable is significant in none of the specifications. The analysis in Table 3 should leave no doubt that our methods are able to detect CAM by identifying a significant relationship

involving the Running mean variable.¹¹ In summary, we are confident that if the DHHC data was consistent with CAM then the methods employed in Table 2 would have detected a relationship between Running mean and Response.

Bias toward the mean across trials

We find evidence that participants are biased toward recently seen lines and not the running mean, which is inconsistent with CAM. However, this is not the unique test of CAM. A benefit of constructing a mathematical model is that it is possible to generate non-obvious predictions that would not be possible without a mathematical model. One non-obvious prediction of CAM relates to the bias toward the mean over the course of the experiment.

CAM holds that participants combine their noisy perception and memory of the target line with their prior beliefs of the distribution of the target lines. Huttenlocher, Hedges, and Vevea (2000, pg. 239) offer 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):

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

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 $g(\cdot)$ of the ratio of these two standard deviations:

$$\lambda = g(\sigma_M / \sigma_P).$$

CAM predicts that the smaller the standard deviation of the prior distribution, the stronger 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 lines, the distribution is unknown and the participant relies on presumably

¹¹ In the Table A13 in the Supplemental Online Appendix, we summarize the analysis with a noise of 50 pixels rather than 25 pixels. This does not change our results.

diffuse priors. However, as the participant repeatedly views target lines of various lengths, the standard deviation of the posteriors decreases. The line lengths that have been seen will have increased posteriors and the line lengths that have not been seen have reduced posteriors. In our setting, lines that are not seen are those shorter than 80 pixels or longer than 368 pixels. This produces a decreasing standard deviation of the prior distribution across trials. Based on this, CAM predicts that the bias toward the mean will increase over the course of the experiment.

We note that this convergence of posteriors occurs regardless of the initial priors. It has been known for some time that, 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).

We use the DHHC data to test this auxiliary prediction of CAM. 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:

$$\text{Running mean bias} = |\text{Target} - \text{Running mean}| - |\text{Response} - \text{Running mean}|.$$

The Running mean bias variable is increasing in the extent to which Response is closer to Running mean than Target is to Running mean.

Over the course of the experiment the participants will learn the distribution with a greater precision, however the rate at which this occurs is not obvious. We therefore offer 5 different specifications. In one specification, the independent variable is simply the trial number. In the second specification, the independent variable is the inverse of the trial number, which we refer to as *Inv. trial*. In the remaining three specifications, we use a categorical variable indicating whether the trial is among the first 5, among the first 10, or among the first 20 trials. If

bias toward the mean is increasing across trials, then the Trial specification would be positive, and the other four specifications would be negative.

As the distribution of targets shifts on trial 96, here we restrict attention to the first 95 trials. Further, because there is not a Running mean that is committed to memory on the first trial, we have a maximum of 94 observations per participant. We perform a random-effects repeated measures analysis, similar to that summarized in Table 2. Finally, because the Running mean bias might depend on the target size and the treatment, we control for this possibility by estimating a dummy variable for each target in both Left and Right Skew treatments. Table 4 summarizes this random-effects analysis.

Table 4: Random-effects regressions of the Running mean bias variable.

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	0.0080 (0.0199)	-10.202 (7.946)	-4.974 [†] (2.721)	-2.418 (1.846)	-1.211 (1.346)
-2 Log L	21644.0	21630.6	21631.0	21633.4	21635.0

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 2 through 95. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 2334 observations. [†] indicates significance at $p < .1$ and * indicates significance at $p < .05$. -2 Log L refers to negative two times the log-likelihood.

In none of the 5 specifications do we find a significant relationship between Running mean and Trial. When we perform the analysis of Table 4, but with fixed-effects, not random-effects, these results are unchanged. These results are not consistent with an auxiliary prediction of CAM.

In the Supplemental Online Appendix (Tables A3-A7) we include additional specifications that examine the bias toward the mean across trials. This includes different measures of the mean bias (the Current mean bias and the Running mean bias expressed as a fraction) and regressions that examine all trials, not just the first half. This produces a total of 30 specifications.

Despite that CAM predicts that bias toward the mean will increase over the course of the experiment, in none of these specifications do we find such a significant relationship. The reader who is concerned that our tests might lack the statistical power to detect an increase in the mean bias across trials should note that 5 of 30 regressions presented either here in the main text or the Supplemental Online Appendix do not even have the same sign as that predicted by CAM.¹²

Responses with zero mass across trials

Now we test another auxiliary prediction of CAM. The model predicts that participants combine their noisy perception and memory with the distribution of the stimuli. This requires that participants learn the distribution across trials. In particular, the Bayesian participant will improve their understanding of the distribution across trials.¹³ This should include learning the lower bound of the distribution and the upper bound of the distribution. Specifically, the Bayesian participant should have diminishing priors on lines that are longer than 368 and shorter than 80, because these lines have zero mass in the probability distribution. Accordingly, the participant should offer such a response with a diminishing frequency across trials.

We define the *Zero mass dummy* to be 1 if the response is greater than 368 or less than 80, and a 0 otherwise. Below, we analyze the first 95 trials. There are 84 instances of a response with zero mass, and 2274 without.

We conduct the analysis similar to that in Table 4, but with two 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

¹² When the analyses are conducted without the treatment-target dummy variables, approximately half of the specifications do not have the same sign as predicted by CAM.

¹³ Again, regardless of the initial prior, Bayesian will learn the true distribution in the long run (Savage, 1954; Blackwell & Dubins, 1962).

dummy variable for every participant. Table 5 summarizes this fixed-effects analysis. We note that CAM would predict a negative estimate for Trial and positive estimates for the others.

Table 5: Fixed-effects logistic regressions of the Zero mass dummy variable.

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	-0.0052 (0.0047)	0.280 (0.995)	0.0369 (0.550)	0.331 (0.391)	0.183 (0.303)
-2 Log L	425.33	426.51	426.59	425.90	426.23

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 1 through 95. We do not provide the estimates of the intercepts, the participant dummy variables, or the treatment-target dummy variables. All regressions have 2358 observations. † indicates significance at $p < .1$. -2 Log L refers to negative two times the log-likelihood.

We do not find a significant relationship in any of the specifications. Also, see Table A11 in the Supplemental Online Appendix for the analysis of every trial rather than simply the first 95. Therefore, we have a total of 10 specifications and in none do we find that zero mass responses are decreasing across trials. This suggests that either the participants are not learning this feature of the distribution or the bias toward the mean is not sufficient to avoid these responses.

Discussion

Experiment 1 was designed to provide evidence that there are no recency effects in serial judgment tasks. However, we find significant recency effects and these are not consistent with CAM. In particular, we find that preceding targets provide a better prediction of the response than the running mean. This becomes more striking when one reflects on the fact that the preceding targets are a limited memory version of the running mean. We also test an auxiliary prediction of CAM that the bias toward the mean will increase across trials. We also do not find evidence of this. Finally, we test a different auxiliary prediction of CAM that there will be a decreasing incidence of responses that are outside of the distribution (larger than the largest line and smaller than the smallest line) as the participants learn the distribution. We do not find

evidence of this either. In contrast to the conclusions of DHHC, we conclude that the judgments in Experiment 1 are not consistent with CAM.

DHHC Experiment 2

Experiment 1 exclusively used asymmetric distributions. Experiment 2 was designed to test whether participants would exhibit a bias toward the running mean with both asymmetric and symmetric distributions that do not vary across trials.

Description of Methods

Participants were asked to judge the same 19 possible target lines as in Experiment 1. These lines were distributed with a Left skew, a Right skew, or a Uniform distribution. Consider again labeling the targets 1 through 19, such that they are increasing length. The Right skew distribution (slightly different from that in Experiment 1) had 19 instances of the target 1, 18 instances target 2, and so on, to 1 instance of target 19. The Left skew distribution (also slightly different from that in Experiment 1) had 19 instances of target 19, 18 instances of target 18, and so on, to 1 instance of target 1. The Uniform distribution had 10 instances of each of the 19 possible target lines. Lines in each of these three distributions were drawn at random without replacement. Therefore, every participant within each treatment estimated the identical set of lines.

Participants were given an initial adjustable line across all trials of either 48 pixels or 400 pixels. Unlike the data associated with Experiment 1, we have access to this information.

The study had 45 participants. Each participant made 190 judgments.¹⁴ Therefore, the total number of judgments was 8550. The reader is referred to DHHC for further details.¹⁵

¹⁴ We note that DHHC reported that they had 36 participants however, the dataset that we have has 45 participants. We note that 10 participants had nonnumeric participant identification codes. It is possible that these were all grouped into a single participant that was recorded as making 1900 judgments. On the other hand, DHHC did not

Repeated measures regressions for preceding target lines

Although the goals of the design of Experiments 1 and 2 are different, our interest in the datasets are the same: to test for the presence of recency effects, whether the mean bias increases across trials, and whether zero mass responses decrease across trials. Therefore, we analyze the dataset using the identical techniques as those used in the analysis of Experiment 1. In order to test for the presence of recency effects, we perform the analysis identical to that summarized in Table 2. Table 6 summarizes this analysis.

Table 6: Random-effects repeated measures regressions of the Response variable.

	No Prec	Prec 1	Prec 3	Prec 5	Prec 10	Prec 15	Prec 20
Target	0.784*** (0.005)	0.784*** (0.005)	0.784*** (0.005)	0.784*** (0.005)	0.783*** (0.005)	0.784*** (0.005)	0.784*** (0.005)
Running mean	0.143*** (0.030)	0.111*** (0.031)	0.0594 [†] (0.032)	0.070* (0.033)	0.105** (0.037)	0.122** (0.040)	0.102** (0.044)
Preceding targets	-	0.026*** (0.005)	0.068*** (0.009)	0.060*** (0.011)	0.031 [†] (0.017)	0.017 (0.023)	0.036 (0.028)
-2 Log L	84304.8	84285.4	84249.2	84284.6	84307.8	84310.0	84308.5

Notes: We provide the coefficient estimates with the standard errors in parentheses. We do not provide the estimates of the intercepts or the covariance parameters. All regressions have 8505 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.

While the evidence is not as stark as that found in Table 2, we find many specifications where the Preceding targets variable is significant. We also find two specifications where the Running Mean variable is not significant at .01.¹⁶ As with the Experiment 1 data, we find we find evidence of recency effects that are not consistent with CAM.¹⁷

Bias toward the mean across trials

report the number of observations. Therefore, we are not able to determine if our dataset is identical to that used in their analysis.

¹⁵ We include this dataset and the code from our analysis in the Supplemental Online Appendix.

¹⁶ See Table A2 in the Supplementary Online Appendix for the fixed-effects version of Table 6. These results are unchanged.

¹⁷ Although Allred et al. (2016) finds that the initial adjustable line affects judgments, when we insert that variable into the regressions summarized in Table 6, we do not find a significant relationship. Despite this, we find a negative correlation between the initial adjustable line and Response ($r(8550) = -0.036, p < .001$).

We also test the auxiliary prediction of CAM that the bias toward the mean will increase across trials. We conduct the analysis using the technique identical to that used in Table 4. Here we consider only the first half of trials so that it is comparable to Table 4. Table 7 summarizes this analysis.

Table 7: Random-effects regressions of the Running mean bias variable.

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	-0.020 (0.016)	-1.259 (6.240)	-0.991 (2.140)	0.140 (1.476)	1.656 (1.079)
-2 Log L	39989.7	39979.3	39981.2	39982.2	39980.5

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 2 through 95. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 4230 observations. † indicates significance at $p < .1$. -2 Log L refers to negative two times the log-likelihood.

In none of the specifications do we find evidence of an increase in Running mean bias across trials. In fact, the First 10 and First 20 specifications have the wrong sign as predicted by CAM. In the Supplemental Online Appendix, we conduct additional analyses that are summarized in Tables A8-A10. This produces a total of 20 specifications. We find 3 significant relationships but we note that they each have the opposite sign as predicted by CAM. Further, 9 of the 20 specifications have coefficient estimates that are opposite sign as predicted by CAM.

Responses with zero mass across trials

In order to learn whether participants exhibit a diminishing incidence of providing a response with a zero mass, we conduct the analysis identical to that in Table 5. Table 8 summarizes this analysis. We note that our data has 181 instances of a response with a zero mass and 4094 without.

Table 8: Fixed-effects logistic regressions of the Zero mass dummy variable.

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	0.0098** (0.0034)	1.735** (0.582)	0.551 (0.356)	-0.0454 (0.297)	-0.527* (0.246)
-2 Log L	870.70	871.68	876.82	879.03	874.14

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 1 through 95. We do not provide the estimates of the intercepts, the participant dummy variables, or the treatment-target dummy variables. All regressions have 4275 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.

We see three significant relationships, however two of them (Trial and First 20) are in the opposite direction as predicted by CAM. In Table A12 in the Supplemental Online Appendix we conduct an additional set of analyses. This produces a total of 10 specifications. There are 4 instances of a significant relationship in the wrong direction and a total of 6 estimates with the wrong sign as predicted by CAM. We conclude that we do not find evidence that the Zero mass dummy is declining across trials.

Conclusions

We have reexamined the data from Experiments 1 and 2 of DHHC. Using their data and their reported technique, we do not find evidence that the running mean is a better predictor of judgments than the recently viewed lines. Further, we perform a different analysis and we find that the participants exhibit a recency bias that is not consistent with CAM.

Further, since the distribution was without replacement, observing a target implies that observing that same target in the future is less likely than if the distribution was with replacement. Despite this experimental design we find evidence of a positive bias toward recent targets when we should actually observe a negative bias toward recent targets. Clearly this reflects even worse on the predictions of CAM.

In order to show that our statistical analysis is capable of detecting judgments that are consistent with CAM, we simulate data that are consistent with a key feature of CAM. Our analysis correctly identifies the simulated data as consistent with CAM. We therefore reject the

criticism that our technique is not capable of accurately detecting a relationship that would be consistent with CAM.

As mathematical models should be used to generate non-obvious and testable predictions, we test two such predictions of CAM. As participants are exposed to stimuli, CAM holds that the participants learn the distribution of the stimuli with greater precision. One prediction is that the bias toward the mean should increase across trials. We do not find evidence of this prediction. Another such prediction of CAM is that, as participants learn the distribution, responses that have zero mass in the distribution (shorter than the minimum or longer than the maximum) should be less frequent across trials. We do not find evidence of this prediction either.

In sum, we do not find evidence consistent with CAM. We have subjected the data from both experiments to (multiple versions) of three tests, and in each we do not find evidence consistent with CAM.

We note that Huttenlocher et al. (2000) and DHHC largely analyze averaged data. The dangers of this have been known for some time (Estes, 1956; Siegler, 1987). In this setting, analyzing averaged data does not permit the investigator to distinguish between the hypothesis that judgments are consistent with CAM and the hypothesis that judgments simply exhibit a bias toward recent stimuli. We also note that these authors do not, as we do, use CAM to generate additional testable hypotheses. While DHHC datasets have failed our tests, we also point out that more work needs to be done before we are able to draw broad conclusions about the merit of CAM. We hope that the methods that we employ will be used to scrutinize other datasets that are considered to be consistent with CAM.

More generally, CAM is a *Bayesian model*. Bayesian models posit that human cognition functions in accordance with Bayes' rule. Specifically, Bayesian models of judgment make the

joint hypothesis that participants learn the distribution of stimuli and they use this information in their judgments. There is an extensive literature on the pros and cons of Bayesian models.¹⁸ A discussion of this literature is beyond the scope of this paper but we note that some authors claim that their results are consistent with Bayesian models (Griffiths & Tenenbaum, 2006; Hemmer & Steyvers, 2009a, 2009b; Lewandowsky, Griffiths, & Kalish, 2009) and others claim that their results are not consistent with Bayesian models (Barth et al., 2015; Cassey et al., 2016; Mozer, Pashler, & Homaei, 2008; Sailor & Antoine, 2005).

To our knowledge, we are the first to find evidence that judgments thought to be consistent with a Bayesian model could be explained by the non-Bayesian use of a set of recent stimuli. Also, to our knowledge, we are the first to apply to Bayesian models of judgment the well-known results that Bayesians with very different initial priors will have posteriors that converge to the true distribution (Savage, 1954; Blackwell & Dubins, 1962). In our analysis we do not see any evidence of learning, either because there was no learning or because the learning did not manifest itself in the judgments. As such, we cannot see how our results are consistent with any Bayesian model of judgment.

Bowers and Davis (2012a) offer a critique of the Bayesian literature and note that authors who tend to claim that their experiments provide evidence in favor of Bayesian models, often do not sufficiently consider non-Bayesian alternatives. By doing this, authors observe judgments that are consistent with the Bayesian model and they conclude that the Bayesian model is supported. By contrast, we directly compare CAM with non-Bayesian explanations by including the Running mean and Previous target variables in the same specifications. Viewing these

¹⁸ See Bowers and Davis (2012a, 2012b), Chater, Tenenbaum, and Yuille (2006), Chater et al. (2011), Elqayam and Evans (2011), Goodman et al. (2015), Griffiths et al. (2012), Hahn (2014), Jones and Love (2011a, 2011b), Marcus and Davis (2013, 2015), Perfors et al. (2011), Petzschner, Glasauer, and Stephan (2015), Tauber et al. (2017), and Tenenbaum, Griffiths, and Kemp (2006).

explanations side-by-side suggests that the non-Bayesian explanation outperforms the Bayesian explanation (CAM).

Any researcher who works on visual judgments or Bayesian models should be concerned with our findings. We acquired datasets that were considered to be consistent with CAM, however careful analysis shows that they are not consistent with CAM. We suspect that our datasets are not unique in this sense and that there exist many such datasets. In fact, given our results, it seems entirely possible that careful analysis of all datasets purportedly offering support to Bayesian models would actually fail to provide evidence supporting these models. We hope that the methods of analysis that we employ will be used to test other Bayesian models of judgment.

It is worth observing that in the past decade psychological science has witnessed a “replication crisis” (Pashler & Harris, 2012; Loken & Gelman, 2017). In the present case, it is now clear that the original analysis presented in DHHC was not correct. Although we do not believe that this error arose from questionable research practices aimed at producing significant results from noise (John, Lowenstein, & Prelec, 2012), but rather it stemmed from a combination of erroneous assumptions about the role of individual differences in the analysis as well as analyses that were in error. While it is difficult to reconstruct the source of the error in an analysis conducted a decade in the past, the present study underscores the importance of maintaining and sharing datasets so that published and unpublished results can be scrutinized. In this spirit, we make our data and our code available in the Supplemental Online Appendix.

We also point out the features of DHHC that increase the chances of arriving at incorrect conclusions. We note that DHHC (as is standard in the psychology literature) presented a statistical analysis with only a single specification. By contrast, we report analyses with multiple

specifications, which vary the explanatory variables, the functional forms, and the assumptions for the error terms. Reporting only a single specification is unhelpful in learning the true nature of complicated phenomena (Simmons, Nelson, & Simonsohn, 2011; Steegen et al., 2016). In any setting, roughly 1 out of 20 specifications will be significant at 5%. If the authors are only expected to report a single specification then it is possible that authors only analyze a single specification and it happens to be the specification that is significant. Additionally, a strategic author could analyze 20 specifications and simply report the one significant specification. However, in our view, if authors were expected to report several specifications then the errors that we find in DHHC would be less likely to go unnoticed. We hope that our reexamination contributes to the ongoing self-reflection on the methods and conventions in the field (Wagenmakers et al., 2011; Wicherts et al., 2016).

Further, we note that DHHC did not report, for instance, the number of observations or their assumptions about the intercepts. Therefore, even though we have the datasets that were used, we cannot be certain that we performed the analyses on the identical set of observations or that we used the identical statistical techniques. It is our view that compelling authors to report these details would be helpful.

Finally, reporting on a different experiment, Sailor and Antoine (2005) find that participants do not make judgments that are consistent with CAM in particular, or Bayesian models in general. It is our view that such evidence is too easily ignored or regarded as a curious anomaly by Bayesian authors. If researchers think that the results of Sailor and Antoine would not replicate then they should test this conjecture. Further, more attention needs to be devoted to settings in which the predictions of any model (and CAM in particular) are violated, rather than

to settings where the predictions are apparently supported. In this way, we will best improve our understanding of how people make judgments.

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.
- 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.

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.

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.

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.

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.

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 23, 178-200.

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.
- 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.
- Hollingworth, H. L. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods, 7*(17), 461-469.
- Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129*, 220-241.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.
- 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.

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.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325), 584-585.

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.

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.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7(6), 531-536.

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.

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

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

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.

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.

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

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.

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.

Xu, J., & Griffiths, T. L. (2010). A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology*, 60(2), 107-126.

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

Preceding targets, fixed-effects analysis: Experiment 1

The analysis summarized in Table 2 finds that the Preceding targets variable offers a better prediction of Response variable than Running mean. However, the reader might be concerned that the results are not robust to the specification of the repeated nature of the data. Below we conduct 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. Table A1 summarizes this fixed-effects analysis.

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

	No Prec	Prec 1	Prec 3	Prec 5	Prec 10	Prec 15	Prec 20
Target	0.813*** (0.005)	0.809*** (0.005)	0.804*** (0.005)	0.803*** (0.005)	0.804*** (0.005)	0.805*** (0.005)	0.806*** (0.005)
Running mean	0.204*** (0.025)	0.109*** (0.027)	0.024 (0.030)	0.018 (0.033)	0.023 (0.037)	0.030 (0.042)	0.038 (0.045)
Preceding targets	-	0.050*** (0.005)	0.095*** (0.009)	0.098*** (0.011)	0.094*** (0.015)	0.089*** (0.017)	0.084*** (0.019)
-2 Log L	45065.8	44990.6	44963.8	44996.6	45030.9	45045.0	45052.7

Notes: We provide the coefficient estimates with the standard errors in parentheses. We do not provide the estimates of the intercepts or participant-specific intercepts. All regressions have 4696 observations. † indicates significance at $p < .1$ and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

Similar to the results of Table 2, here we find that the Preceding targets variable is significant in every specification. Further, in all but the first two specifications, the Running mean variable is not significant.

Preceding targets, fixed-effects analysis: Experiment 2

Table A2 summarizes the fixed-effects analysis version of Table 6.

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

	No Prec	Prec 1	Prec 3	Prec 5	Prec 10	Prec 15	Prec 20
Target	0.783*** (0.005)	0.783*** (0.005)	0.783*** (0.005)	0.783*** (0.005)	0.783*** (0.005)	0.783*** (0.005)	0.783*** (0.005)
Running mean	0.128*** (0.034)	0.095** (0.035)	0.039 (0.036)	0.050 (0.037)	0.087* (0.041)	0.105* (0.045)	0.082 [†] (0.048)
Preceding targets	-	0.026*** (0.005)	0.069*** (0.009)	0.060*** (0.011)	0.032 [†] (0.017)	0.019 (0.023)	0.038 (0.028)
-2 Log L	83923.2	83903.5	83867.0	83902.4	83926.0	83928.2	83926.6

Notes: We provide the coefficient estimates with the standard errors in parentheses. We do not provide the estimates of the intercepts or participant-specific intercepts. All regressions have 8505 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.

We find 5 specifications where the Running mean is not significant at .01. We also find 3 specifications where the preceding targets variable is significant at .01. In summary, we find significant recency effects that are not consistent with CAM.

Bias towards the mean across trials: Experiment 1

In order to verify the robustness of the analysis summarized in Table 4, here we perform a nearly identical set of regressions. One feature of Experiment 1 is that the mean of the distribution switched. Therefore, we run a specification with *Current mean bias*, rather than the Running mean bias, as the dependent variable. Table A3 summarizes this analysis.

Table A3: Random-effects regressions of the Current mean bias variable.

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	0.017 (0.020)	-12.482 (7.838)	-5.256 [†] (2.684)	-3.392 [†] (1.820)	-1.878 (1.328)
-2 Log L	21579.8	21566.1	21566.9	21568.0	21570.1

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 2 through 95. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 2334 observations. [†] indicates significance at $p < .1$ and * indicates significance at $p < .05$. -2 Log L refers to negative two times the log-likelihood.

Similar to the analysis summarized in Table 4, here we do not find evidence of a significant increase in the Current mean bias over trials. We also note that these results are unchanged when the regressions are performed with fixed-effects, rather than random-effects.

In Tables 4 and A3 we respectively examined the Running mean bias and the Current mean bias across trials in the first half of the experiment. Here we examine data from both halves of the experiment. However, since there was a change in the distribution in trial 96, we employ a variable that accounts for this change. We define the *Round* variable to be the number of trials that the participant had been exposed to the particular distribution. In other words, the Round variable and the Trial variable are identical for trials less than 96, and the Round variable is the Trial variable minus 95 for trials greater than or equal to 96. We have constructed the analogous 5 independent variables but for Rounds, not Trials. Table A4 summarizes the regressions of the Running mean bias across Rounds.

Table A4: Random-effects regressions of the Running mean bias variable.

	Round	Inv. Round	First 5	First 10	First 20
Round	-0.0283 [†] (0.0153)	3.057 (4.272)	-1.277 (1.978)	0.229 (1.391)	0.851 (1.031)
-2 Log L	44644.6	44636.3	44637.9	44639.0	44639.0

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 2 through 190. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 4696 observations. [†] indicates significance at $p < .1$ and * indicates significance at $p < .05$. -2 Log L refers to negative two times the log-likelihood.

Similar to the results summarized in Table 4, none of the specifications are significant at .05. Additionally, a fixed-effects specification does not change the results. We also note that the coefficient estimates in the Round, Inv. Round, First 10, and First 20 have the opposite signs as predicted by CAM. Table A5 summarizes regressions of Current mean bias across Rounds.

Table A5: Random-effects regressions of the Current mean bias variable.

	Round	Inv. Round	First 5	First 10	First 20
Round	-0.018 (0.014)	-3.799 (3.986)	-1.424 (1.846)	-0.871 (1.298)	-0.567 (0.962)
-2 Log L	44002.5	43991.9	43993.8	43994.6	43995.3

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 2 through 190. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 4696 observations. † indicates significance at $p < .1$. -2 Log L refers to negative two times the log-likelihood.

Again, we see that none of the specifications are significant. We also note that the fixed-effects specification does not change the results. We also note that the Round estimate has the opposite sign as predicted by CAM.

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:

Running mean bias ratio =

$$| \text{Target} - \text{Running mean} | / [| \text{Target} - \text{Running mean} | + | \text{Response} - \text{Running mean} |]$$

Below we use the Running mean bias ratio as the dependent variable. Table A6 summarizes this random-effects analysis on the first half of trials.

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

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	0.00004 (0.00013)	-0.0631 (0.0498)	-0.0201 (0.0171)	-0.0091 (0.0116)	-0.0070 (0.0084)
-2 Log L	1649.3	1662.8	1660.4	1658.9	1658.3

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 2 through 95. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 2334 observations. † indicates significance at $p < .1$. -2 Log L refers to negative two times the log-likelihood.

Although the sign of each estimate is that predicted by CAM, we also note that none are significant. Table A7 summarizes the analysis, similar to Table A6, but on every trial.

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

	Round	Inv. Round	First 5	First 10	First 20
Round	0.00001 (0.00009)	-0.0162 (0.0247)	-0.0151 (0.0114)	-0.0062 (0.0080)	-0.0027 (0.0060)
-2 Log L	3368.5	3380.2	3380.0	3378.1	3377.1

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 2 through 190. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 4696 observations. † indicates significance at $p < .1$. -2 Log L refers to negative two times the log-likelihood.

Similar to that found in Table A6, none of the estimates are significant.

Bias towards the mean across trials: Experiment 2

Whereas Table 7 analyzed the Running mean bias in Experiment 2 for the first half of the trials, Table A8 summarizes our random-effects analysis across all trials.

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

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	-0.0149** (0.006)	3.077 (5.973)	-0.457 (2.167)	0.941 (1.468)	2.286* (1.039)
-2 Log L	81099.5	81092.1	81094.3	81094.7	81091.0

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 2 through 190. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 8505 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.

Here we see two specifications where the Running mean bias significantly decreases across trials (Trial and First 20). However, in no specification does the Running mean bias significantly increase across trials. We also note that the signs of each of these estimates, with exception of First 5, are the opposite of that predicted by CAM. Finally, a fixed-effects specification does not change the result that the running bias does not increase across trials.

Table A9 summarizes the analysis of the Running mean bias ratio variable across the first half of trials.

Table A9: Random-effects regressions of the Running mean bias ratio variable.

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	-0.00002 (0.00009)	-0.0677 [†] (0.0348)	-0.0152 (0.0119)	-0.00727 (0.00824)	0.00320 (0.00603)
-2 Log L	3327.5	3343.2	3338.9	3337.3	3336.2

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 2 through 95. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 4230 observations. [†] indicates significance at $p < .1$ and * indicates significance at $p < .05$. -2 Log L refers to negative two times the log-likelihood.

In none of the specifications do we detect a significant relationship. In fact, the Trial and the First 20 specifications are the opposite sign as that predicted by CAM. Below we conduct the analysis of Table A9 on each trial. Table A10 summarizes this analysis. Note that 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 A10 than in Table A8

Table A10: Random-effects regressions of the Running mean bias ratio variable.

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	-0.00008* (0.00003)	-0.0329 (0.0323)	-0.010 (0.0117)	-0.0036 (0.0079)	0.00685 (0.00562)
-2 Log L	7112.5	7120.1	7117.7	7116.4	7117.0

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 2 through 190. We do not provide the estimates of the intercepts, the covariance parameters, or the treatment-target dummy variables. All regressions have 8504 observations. [†] indicates significance at $p < .1$, * indicates significance at $p < .05$, and ** indicates significance at $p < .01$. -2 Log L refers to negative two times the log-likelihood.

Here we detect a single significant relationship (Trial) and it is in the opposite direction as predicted by CAM. Further we note that only 3 of the 5 estimates even have the same sign as that predicted by CAM.

Responses with zero mass across trials: Experiment 1

Table 5 analyzed the Zero mass dummy variable across the first half of trials. Table A11 summarizes the regressions of the Zero mass across all trials. We note that there are 158 observations with a zero mass and 4562 without.

Table A11: Random-effects logistic regressions of the Zero mass dummy variable.

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	-0.0019 (0.0018)	0.763 (0.935)	0.176 (0.525)	0.433 (0.371)	0.273 (0.282)
-2 Log L	868.37	868.89	869.37	868.21	868.58

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 1 through 190. We do not provide the estimates of the intercepts, the participant dummy variables, or the treatment-target dummy variables. All regressions have 4720 observations. † indicates significance at $p < .1$. -2 Log L refers to negative two times the log-likelihood.

Similar to the results summarized in Table 5, none of the specifications are significant.

Responses with zero mass across trials: Experiment 2

Whereas Table 8 analyzed the Zero mass dummy in Experiment 2 for the first half of the trials, Table A12 summarizes our random-effects analysis across all trials. We note that there are 384 observations with a zero mass and 8166 without.

Table A12: Fixed-effects logistic regressions of the Zero mass dummy variable.

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	0.0026* (0.0012)	1.277* (0.561)	0.467 (0.342)	-0.124 (0.289)	-0.600* (0.235)
-2 Log L	1854.94	1855.61	1858.33	1859.88	1852.85

Notes: We provide the coefficient estimates with the standard errors in parentheses. We restrict attention to trials 1 through 190. We do not provide the estimates of the intercepts, the participant dummy variables, or the treatment-target dummy variables. All regressions have 8550 observations. † indicates significance at $p < .1$, * indicates significance at $p < .05$, and ** indicates significance at $p < .01$. -2 Log L refers to negative two times the log-likelihood.

Here again, the results are reminiscent of those in Table 5. The Trial and First 20 variables are significant, albeit in the opposite direction as predicted by CAM. However, the Inv. Trial variable is significant and in the direction as predicted by CAM.

Simulated Response50 variable: Experiment 1

In Table 3 we analyzed the simulated Response25 variable. Here we perform the identical analysis with the simulated Response50 variable, which contains noise with a standard deviation of 50 pixels, rather than 25 pixels.

Table A13: Random-effects repeated measures regressions of the simulated Response50 variable.

	No Prec	Prec 1	Prec 3	Prec 5	Prec 10	Prec 15	Prec 20
Target	0.885*** (0.008)	0.886*** (0.008)	0.886*** (0.008)	0.885*** (0.008)	0.885*** (0.008)	0.884*** (0.008)	0.884*** (0.008)
Running mean	0.113*** (0.022)	0.117*** (0.023)	0.118*** (0.023)	0.114*** (0.024)	0.113*** (0.025)	0.111*** (0.026)	0.107*** (0.027)
Preceding targets	-	-0.006 (0.008)	-0.008 (0.012)	-0.002 (0.014)	-0.0002 (0.017)	0.003 (0.018)	0.007 (0.019)
-2 Log L	49409.0	49416.2	49415.5	49415.6	49415.3	49415.1	49414.9

Notes: We provide the coefficient estimates with the standard errors in parentheses. We do not provide the estimates of the intercepts or the covariance parameters. All regressions have 4696 observations. † indicates significance at $p < .1$ and *** indicates significance at $p < .001$. -2 Log L refers to negative two times the log-likelihood.

Despite a different noise component than in Table 3, in every specification we find a significant relationship between Running mean and the Response50 variable. Again, there should be no doubt that our analysis is capable of detecting a relationship that is consistent with CAM.

We note that the noise in the analysis of Table A13 exceeds that in our original analysis, as can be seen by comparing the -2 Log L values. We also note that the noise in the analysis of Table 3 is less than that in the analysis of Table 2, as can be seen by comparing the -2 Log L values. Given the results of Tables 3 and A13, we reject the criticism that the declining standard deviation of Running mean prevents satisfactory estimates of the coefficient of the Running mean variable.