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

Duffy, Sean and Smith, John

Rutgers University-Camden

6 January 2017

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

January 6, 2017

*We thank Roberto Barbera, I-Ming Chiu, L. Elizabeth Crawford, and Johanna Hertel 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. CAM predicts that there will exist a bias toward the running mean of the lines and that judgments will not be differentially affected by recent stimuli. The authors report that their analysis provides evidence consistent with CAM. 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 a test of an auxiliary prediction of CAM: that the bias towards the mean will increase with exposure to the distribution. We do not find such a relationship. Fourth, we produce a simulated dataset that is consistent with 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.

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

A well-known experimental effect is that participants tend to have 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, and Vevea, 2000) holds that participants imperfectly remember stimulus and they improve accuracy by considering information about the probability distribution of the stimuli. In particular, CAM suggests that participants optimally combine the distribution and their imperfect memory of the stimuli by employing Bayes' Rule. Whereas such judgments minimize the error resulting from the imperfect memory of the stimulus, 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.

Duffy, Huttenlocher, Hedges, and Crawford (2010), hereafter referred to as DHHC, studied whether judgments of the lengths of lines are consistent with CAM.

DHHC Experiment 1

Experiment 1 was designed to test whether participants have a bias toward the running mean of stimuli or toward recent values of the 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 Left skew distribution (long lines less likely than short lines) 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 Right 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, 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.

Description of the dataset

Among these 4750 observations, there are 30 missing values for the response line. It seems as if the 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. (2017) 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 Left skew then Right 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.

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 that CAM is consistent with their results.

Our reexamination

Before we begin with our reexamination, we say a few words about the methods employed by DHHC. The dataset that we obtained 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.

Although we are not certain, it seems from the description of the analysis that DHHC estimated a specification in which the intercept was assumed to be zero. However, the authors did not justify this assumption. Therefore in our reexamination, 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 specifications in Table 1.

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

		Preceding 20 targets p-values					
		$p \geq .1$	$.1 > p \geq .05$	$.05 > p \geq .01$	$.01 > p \geq .001$	$.001 > p$	Total
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 9 of the regressions with an assumed zero intercept and it is significant in 10 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 significant in 11 of the regressions that assume a zero intercept and it is not significant in 23 regressions where the intercept is not constrained to be zero.

We admit that an error in the execution of the analysis is likely responsible for the results presented by DHHC. However, these erroneous results drastically affect the conclusions. 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, it is possible that there is not enough statistical power in the participant-level regressions to find a significant relationship. Further,

their analysis does not provide an aggregate estimate of the relationships among the variables. Finally, running participant-level regressions renders the summary of the analyses to be needlessly cumbersome.

Here we employ standard repeated measures techniques in order to remedy these shortcomings. We run many different specifications that include different numbers of preceding targets. 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*. We also 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.[1]

[1] 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***	0.808***	0.803***	0.802***	0.804***	0.805***	0.807***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Running mean	0.200***	0.115***	0.040	0.037	0.046	0.056	0.065
	(0.024)	(0.025)	(0.028)	(0.030)	(0.034)	(0.037)	(0.040)
Preceding targets	-	0.0495***	0.0925***	0.094***	0.088***	0.081***	0.075***
		(0.005)	(0.009)	(0.011)	(0.014)	(0.016)	(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 $p < .1$, * 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 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.[2] This suggests that Preceding targets tend to be a better predictor of Response than Running mean.[3]

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 question, we simulated a very simple dataset that is consistent with CAM and has parameters similar to that found in the DHHC data. We took the sequence of Target lines from Experiment 1 and added normally

[2] 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.
[3] 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.

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 the Memory and the 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 CAM in that Response25 is biased toward Running mean but not toward recent lines. 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***	0.899***	0.898***	0.898***	0.897***	0.898***	0.899***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Running mean	0.105***	0.106***	0.105***	0.104***	0.103***	0.104***	0.105***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)
Preceding targets	-	-0.003	0.002	0.004	0.006	0.004	0.001
		(0.004)	(0.006)	(0.007)	(0.008)	(0.008)	(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 $p < .1$, * 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 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 noting a significant relationship involving

the Running mean variable.[4] 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 observations

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(.) 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

[4] In the Table A7 in the Supplemental Online Appendix, we perform the analysis with a noise of 50 pixels rather than 25 pixels. This does not change our results.

diffuse prior beliefs. 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, the lines that are not seen are those shorter than 80 pixels and longer than 368 pixels, and 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 use the DHHC data to test this auxiliary prediction of CAM. In order to test this, 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. 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	-0.0056	-5.448	-3.354	-2.267	0.293
	(0.022)	(8.729)	(2.981)	(2.030)	(1.479)
-2 Log L	22315.0	22302.7	22304.0	22304.8	22306.6

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 or the covariance parameters. All regressions have 2334 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 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. We also note that in the Trial and the First 20 specifications, the sign indicates that bias toward the mean is actually decreasing over trials. These results are not consistent with an auxiliary prediction of CAM.[5]

Despite that CAM predicts that bias toward the mean will increase over the course of the experiment, we do not find evidence of this. This seems to provide further evidence against CAM. 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 10 of 20 random-effects regressions presented either in the main text or the Supplemental Online Appendix do not even have the same sign as that predicted by CAM.[6]

Discussion

[5] See Tables A3, A4, and A5 in the Supplemental Online Appendix for additional specifications.
[6] Additionally, 10 of the 20 analogous and unreported fixed-effects regressions do not have the signs as predicted by CAM.

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. The judgments in Experiment 1 are not consistent with CAM.

DHHC Experiment 2

Experiment 2 was designed to test whether participants have a bias toward the center of the computer monitor.

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 Left 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 Right 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 in 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. Therefore, the total number of judgments was 8550. The reader is referred to DHHC for further details.

Description of the dataset

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

Repeated measures regressions for preceding target lines

Although the goals of Experiments 1 and 2 are different, our interest in the dataset is the same: to test for the presence of recency effects and whether the mean bias increases 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 5 summarizes this analysis.

Table 5: 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***	0.784***	0.784***	0.784***	0.783***	0.784***	0.784***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Running mean	0.143***	0.111***	0.0594†	0.070*	0.105**	0.122**	0.102**
	(0.030)	(0.031)	(0.032)	(0.033)	(0.037)	(0.040)	(0.044)
Preceding targets	-	0.026***	0.068***	0.060***	0.031†	0.017	0.036
		(0.005)	(0.009)	(0.011)	(0.017)	(0.023)	(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.[7] As with the Experiment 1 data, we find we find evidence of recency effects that are not consistent with CAM.[8]

Bias toward the mean across observations

We also test an 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 95 trials so that it is comparable to Table 4. Table 6 summarizes this analysis.

Table 6: Random-effects regressions of the Running mean bias variable.

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	-0.032†	1.208	-0.583	1.154	2.686*
	(0.018)	(6.846)	(2.353)	(1.614)	(1.182)
-2 Log L	41125.9	41117.2	41119.3	41119.6	41115.6

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 or the covariance parameters. All regressions have 4230 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 none of the specifications do we find evidence of an increase in Running mean bias across trials. In fact, in contrast to the prediction of CAM, we see significantly more Running mean bias in the first 20 trials than in later trials. Also in the fixed-effects specifications we do not find evidence of the mean bias increasing across trials.[9]

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

[7] See Table A2 in the Supplementary Online Appendix for the fixed-effects version of Table 5. These results are similar.

[8] Although Allred et al. (2017) finds that the initial adjustable line affects judgments, when we insert that variable into the regressions summarized in Table 5, we do not find a significant relationship. Despite this, we find a negative correlation between the initial start line and Response (r (8550) = -0.036, $p < .001$).

[9] In Table A6 of the Supplemental Online Appendix we analyze the data across all trials. There we find two significant relationships that are the opposite of that predicted by CAM, yet we do not find a significant relationship that is predicted by CAM.

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.

We also tested an auxiliary prediction of CAM. As participants are exposed to stimuli, they learn the distribution of the stimuli with greater precision, and CAM predicts that they use more of this information in their judgments. We do not find evidence consistent with this prediction.

In order to show that our statistical analysis is capable of detecting judgments that are consistent with CAM, we simulate data that is consistent with 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. While our employed statistical techniques are perhaps not familiar to every psychologist, we emphasize how standard they are in other fields.

Bowers and Davis (2012) offer a critique of the Bayesian literature and note that authors that 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 confirmed. 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 explanations side-by-side suggests to us that the non-Bayesian explanation outperforms the Bayesian explanation (CAM). Our results suggest that more thorough analysis should have been conducted before DHHC concluded that, “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 towards a category’s central value.”

Any researcher who works on visual judgments should be concerned with our findings. We acquired a dataset that was considered to be consistent with CAM, however careful analysis shows that it is not consistent with CAM. We suspect that this dataset is 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 of Bayesian models would actually fail to provide evidence supporting these models. We encourage authors, who possess datasets on the topic of Bayesian judgments, to send them to the corresponding author of this article so that they can be analyzed by the techniques used in this paper.

We also note that DHHC (as is standard in the psychology literature) presented a statistical analysis with only a single specification. In our view, this is unhelpful in learning the true nature of complicated phenomena. 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 perform a single specification and it happens to be the specification that is significant. Additionally, a strategic author could perform 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.

Further, we note that the DHHC did not report the number of observations in their analyses. Therefore, even though we have the datasets that were used, we cannot be certain if we performed the analyses on the identical set of observations. Compelling authors to report the number of observations would avoid this problem.

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 confirmed. In this way, we will best improve our understanding of how people make judgments.

References

Allred, S., Crawford, L.E., Duffy, S., & Smith, J. (2017). Working memory and spatial judgments: Cognitive load increases the central tendency bias. *Psychonomic Bulletin & Review*, forthcoming.

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

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.

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

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.

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

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 the Response variable than the 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***	0.809***	0.804***	0.803***	0.804***	0.805***	0.806***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Running mean	0.204***	0.109***	0.024	0.018	0.023	0.030	0.038
	(0.025)	(0.027)	(0.030)	(0.033)	(0.037)	(0.042)	(0.045)
Preceding targets	-	0.050***	0.095***	0.098***	0.094***	0.089***	0.084***
		(0.005)	(0.009)	(0.011)	(0.015)	(0.017)	(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 $p < .1$, * 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.

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 5.

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***	0.783***	0.783***	0.783***	0.783***	0.783***	0.783***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Running mean	0.128***	0.095**	0.039	0.050	0.087*	0.105*	0.082†
	(0.034)	(0.035)	(0.036)	(0.037)	(0.041)	(0.045)	(0.048)
Preceding targets	-	0.026***	0.069***	0.060***	0.032†	0.019	0.038
		(0.005)	(0.009)	(0.011)	(0.017)	(0.023)	(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 partcipant-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 observations: 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	0.00311	-7.217	-3.497	-3.134	-0.304
	(0.0219)	(8.724)	(2.980)	(2.029)	(1.478)
-2 Log L	22310.7	22298.1	22299.6	22299.3	22302.3

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 or the covariance parameters. All regressions have 2334 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.

Similar to the analysis summarized in Table 4, here we do not find evidence of an 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	-0.026†	-0.820	0.120	-0.609	0.142
	(0.016)	(4.432)	(2.050)	(1.444)	(1.070)
-2 Log L	45200.1	45191.6	45193.1	45193.7	45194.4

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 or the covariance parameters. All regressions have 4696 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.

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, First 5 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	-0.031†	5.157	0.076	0.250	1.395
	(0.016)	(4.496)	(2.080)	(1.465)	(1.086)
-2 Log L	45338.3	45329.5	45332.3	45333.0	45332.0

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 or the covariance parameters. All regressions have 4696 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, we see that none of the specifications are significant at .05. We also note that the fixed-effects specification does not change the results. Although none of the specifications are significant at .05, we note that the Round specification is significant at .1. However, each of these estimates have the opposite sign as predicted by CAM.

Bias towards the mean across observations: Experiment 2

Whereas Table 6 analyzed the Running mean bias in Experiment 2 for the first half of the trials, Table A6 summarizes our random-effects analysis across all trials.

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

	Trial	Inv. Trial	First 5	First 10	First 20
Trial	-0.020***	6.935	0.312	2.002	3.343**
	(0.006)	(6.626)	(2.407)	(1.626)	(1.151)
-2 Log L	83182.6	83177.2	83180.3	83179.5	83173.3

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 or the covariance parameters. All regressions have 8550 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 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.

Simulated Response50 variable: Experiment 1

In Table 3 we analyzed the simulated Respone25 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 A7: 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***	0.886***	0.886***	0.885***	0.885***	0.884***	0.884***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Running mean	0.113***	0.117***	0.118***	0.114***	0.113***	0.111***	0.107***
	(0.022)	(0.023)	(0.023)	(0.024)	(0.025)	(0.026)	(0.027)
Preceding targets	-	-0.006	-0.008	-0.002	-0.0002	0.003	0.007
		(0.008)	(0.012)	(0.014)	(0.017)	(0.018)	(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 $p < .1$, * 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.

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 A7 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 A7, we reject the criticism that the declining standard deviation of Running mean prevents satisfactory estimates of the coefficient of the Running mean variable.