Psychonomic Bulletin & Review (2019) 26:693–698
https://doi.org/10.3758/s13423-019-01592-y

Reply to Duffy and Smith's (2018) reexamination

L. Elizabeth Crawford[1]

Published online: 22 March 2019

Abstract
Duffy, Huttenlocher, Hedges, and Crawford (2010, *Psychonomic Bulletin & Review*, *17*[2], 224–230) examined whether the well-established central tendency bias in people's reproductions of stimuli reflects bias toward the mean of an entire presented distribution or bias toward only recently seen stimuli. They reported evidence that responses were biased toward the long-run mean and found no evidence that they were biased toward the most recent stimuli. Duffy and Smith (2018) reexamine the data using a different analytical strategy and argue that estimates are biased by recent stimuli rather than toward the long-run mean. I argue that this reanalysis misses a true effect of the running mean and that the data are (mostly) consistent with the claims in the original work. I suggest that these results, and many other null results presented by Duffy and Smith, do not have major theoretical significance for the category adjustment model and similar Bayesian models. (Code and data available: https://osf.io/tkqvn.)

Keywords Human memory · Statistical inference · Categorization

At issue in earlier work by Duffy, Huttenlocher, Hedges, and Crawford (2010, hereafter called DHHC) was whether the well-established central tendency bias in people's reproductions of stimuli reflects bias toward the mean of an entire distribution or bias toward recently seen stimuli. DHHC reported evidence that responses were biased toward the long-run mean rather than toward the most recent stimuli and presented these results as consistent with the category adjustment model (CAM), a Bayesian model of stimulus estimation (Huttenlocher, Hedges, & Vevea, 2000). In their reexamination, Duffy and Smith (DS; 2018) point out some flaws in the original work, apply a different analytical approach to the data, and conclude that estimates are biased toward recent stimuli rather than toward the running mean. They interpret their findings as a refutation of CAM, and of Bayesian models more generally. Here, I argue that the DS failure to detect the impact of the running mean on estimates is likely a Type II error. I also suggest that the paper overstates the theoretical significance of its findings and of the original findings reported in DHHC, and that it presents a distorted view of CAM.

✉ L. Elizabeth Crawford
lcrawfor@richmond.edu

[1] University of Richmond, Richmond, VA, USA

An important contribution of CAM was that it provided a rational account of the central tendency bias in people's reproductions of stimuli. CAM modeled responses as being drawn from the posterior distribution created by a Bayesian combination of two Gaussians—one representing the unbiased but noisy memory trace of the stimulus and the other representing prior knowledge about the category from which it was drawn (i.e., the generating distribution). The resulting responses are a weighted average of the trace memory mean and the mean of the prior, and the weighting is determined by the relative precision of the trace and the prior:

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

Where ρ: mean of the prior
M: mean of the memory trace.

This combination produces responses that are biased toward the mean of the prior. Central to the present argument, CAM specifically predicts that the degree of that bias will depend on *the deviation of M from* ρ. For stimuli that vary in size, to the degree that stimuli are larger than ρ, they will be underestimated; to the degree that they are smaller than ρ, they will be overestimated, and stimuli in the middle (i.e., where M = ρ) will produce unbiased responses. Thus the normal–normal combination on which CAM is based predicts a linear relationship between bias (i.e., response–stimulus) and the deviation of M from ρ.

DHHC asked, where is ρ? That is, toward what point do people bias their estimates? Earlier work has

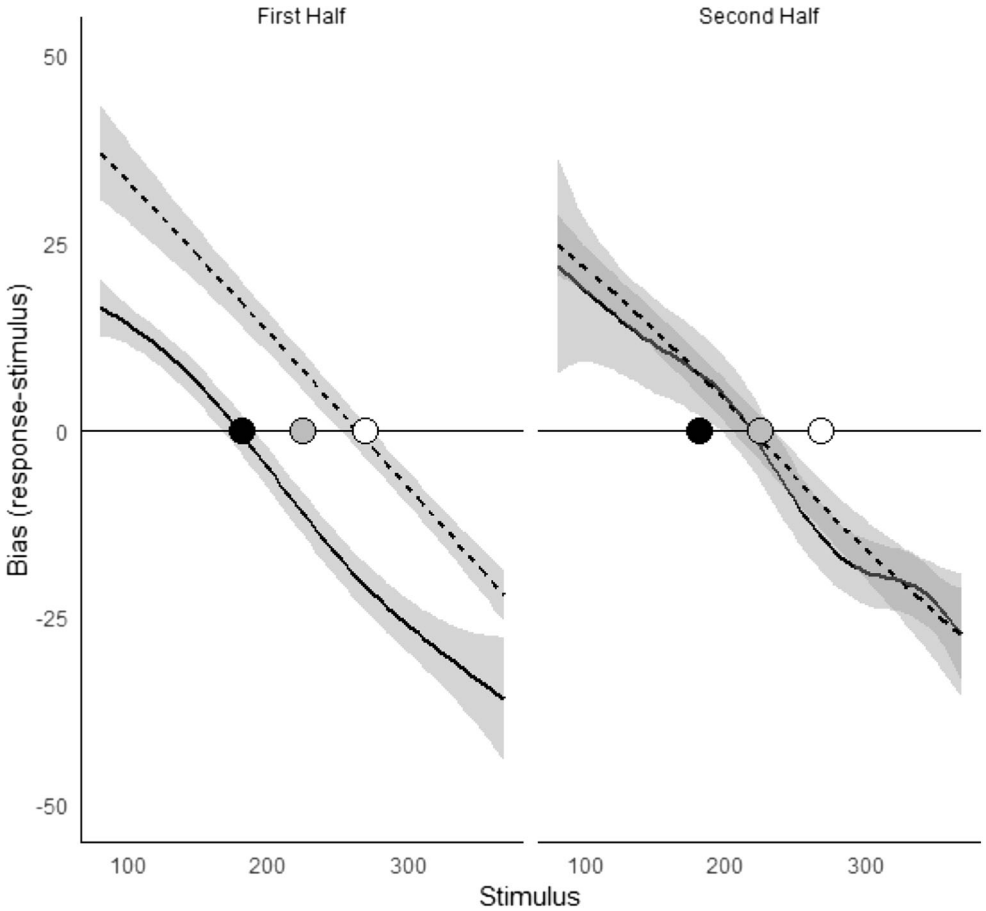

Fig. 1 Bias (i.e., response–stimulus) by stimulus value for each condition in the first half (left panel) and second half (right panel) of DHHC Experiment 1. The first 20 trials of the second half are excluded in order to give the later trials a better chance of being unaffected by the running mean. Dotted lines indicate the condition that first viewed longer lines; solid lines indicate the condition that first viewed shorter ones. White circle indicates the mean of the longer distribution, black one the mean of the shorter distribution, and gray one the mean of all stimuli. Gray bands here and in subsequent figures represent 95% confidence intervals calculated in ggplot2 using the general additive model method (Wickham, 2009)

suggested that, at least in some contexts, participants use a prior that corresponds to (or at least tracks variation in) the mean and variability of the generating distribution (e.g., Duffy, Huttenlocher, & Crawford, 2006; Huttenlocher et al., 2000). Others noted that the central tendency bias could also be achieved if people shift estimates toward only the previous instance (or few instances) without taking into account the entire distribution (Choplin & Hummel, 2002). DHHC (Experiment 1) sought to distinguish these possibilities by having participants view and reproduce line lengths and using a distribution that changed between the first half of trials and the second half so that the long-run mean and short-run mean would differ. Both distributions had the same range, but they were skewed so that in the first half, the distribution's average was 180 pixels for some participants and 268 pixels for others, and in the second half, the distributions flipped. The resulting bias in estimates is shown in Fig. 1.

During the first half of trials (left panel), responses are biased toward the mean of the first-half distribution. This is shown by the negative slope (smaller stimuli are overestimated, larger stimuli underestimated), and by the differing intercepts (because each condition was biasing toward a different central value). During the first half of trials, within each condition the cumulative mean and the mean of recent stimuli are the same (on average), so a process that adjusted estimates toward either of these would produce the same overall pattern of bias. In the second half of trials, the cumulative mean and the recent mean diverge. If recent stimuli are most influential, estimates should now shift toward the mean of the second-half distribution, which is different for the two conditions and so should again produce two separate bias curves with different intercepts. However, if the cumulative mean is most influential, the two conditions should converge as participants in both conditions integrate their exposure to both distributions and acquire the same cumulative mean.

Figure 1 shows that during the second half of trials, responses are biased toward the long-run mean rather than the short-run mean.[1] The pattern seems incompatible with the DS paper's conclusion that estimates are biased toward recent stimuli rather than the long-run mean. That conclusion is based on a null result from an analytical approach that can be shown to be inadequately sensitive to the effect of the running mean on estimates that CAM predicts.

In order to assess the impact of the running mean and the mean of *N* recent stimuli on responses, DS fit several regressions of the form:

$$\text{Response} = \beta_0 + \beta_1\text{Stimulus} + \beta_2\text{RunningMean} + \beta_3\text{Last}N\text{Mean} + \varepsilon \quad (2)$$

This approach assesses whether variation in the running mean and mean of recent stimuli independently predict variation in responses, but does not directly assess whether estimates are biased toward the running mean in the manner that CAM predicts. By treating each predictor as independent, it does not capture the constrained relation between these predictors in CAM (i.e., that the weight given to the prior in estimates is inversely proportional to the weight given to the stimulus).

One way to model this aspect of CAM is to use regression through the origin (RTO) by dropping the intercept term from the DS analysis. This was done in the DHHC paper, albeit without explanation. Although RTO is controversial, it is appropriate to test a model in which the response depends on predictors that are proportionally related (Cornell, 2011; Weisberg, 1985), as is the case in CAM. An alternative and more intuitive approach is to model CAM's prediction that bias in responses depends on the deviation of a stimulus from the prior. If the prior toward which people bias their estimates is the running mean, and the trace memory of a stimulus is centered on the stimulus value, then Equation 1 can be revised to:

$$\text{Response} = (\lambda)\ \text{Running mean} + (1-\lambda)\text{Stimulus}. \quad (3)$$

This is equivalent to:

$$\text{Response}-\text{Stimulus} = -\lambda\ (\text{Stimulus}-\text{Running mean}). \quad (4)$$

To test whether estimates are biased toward the running mean without using RTO, one can run regressions of this form:

$$\text{Bias} = \beta_{0+}\ \beta_1\text{RunMeanDeviation} + \varepsilon, \quad (5)$$

Where Bias*:* response–stimulus

RunMeanDeviation: stimulus-running mean.

CAM predicts that β_1 will be a negative value, the extremity of which estimates the weight given to the running mean (i.e., $-\beta_{1\ =}\ \lambda$) and that β_0 will be zero. To test whether estimates are also biased toward the mean of *N* recent stimuli, one can also include each stimulus's deviation from that mean:

$$\text{Bias} = \beta_{0+}\ \beta_1\text{RunMeanDeviation} + \beta_2\text{LastNMeanDeviation} + \varepsilon. \quad (6)$$

This approach, hereafter called "deviations analysis," produces different results from the DS analysis. To see why, consider a simulated experiment in which participants view and reproduce stimuli drawn at random without replacement from a set of lines ranging from 80 to 368 in equal increments of 16, each shown 10 times (similar to the uniform condition of Huttenlocher et al., 2000). Here, responses were generated to include both bias toward the running mean (i.e., mean of all previous stimuli, not including the present trial) and toward the mean of the three stimuli that preceded the present trial and some noise:

$$\text{Response} = \text{Stimulus} + .8 \times \text{Stimulus} + .1 \times \text{RunningMean} + .1 \times \text{Last3Mean} + \text{Noise}. \quad (7)$$

From this simulation, we can calculate the running mean for each participant at each trial number to see how it changes across trials, as shown in Fig. 2.

The variability in the running mean decreases across trials, as it must with increasing sample size.[2] The variability of the mean of the last three (or *n*) stimuli does not suffer the same fate because it is always based on only three (or *n*) data points (see Fig. 3).

Because the DS analysis treats the running mean as an independent predictor of responses, its sensitivity to the effect of the running mean will depend on the variability in the running mean, which decreases across trials. It will detect the effect of the running mean in initial trials, but not later trials. Analyzing each half of this simulated data set with the DS analyses and with the deviations analysis produces the estimates in Table 1.

In the first half of trials (left column of Table 1), both analyses reveal a statistically significant effect of the running mean and the last-three mean, although the standard errors in the DS analysis show that it estimates the running mean effect less precisely than it does the effect of the last-three mean and also less precisely than does the deviations analysis. In the second half (right column), where there is little variation in

[1] DHHC calculated means and 95% confidence intervals for the point of zero bias (i.e., the value of *x* where *y* = 0) in each condition and each half of the experiment. They reported that in the first half, the 95% CIs included the mean of the distribution shown in that half, but in the second half, the 95% CIs were shifted toward the cumulative mean and did not include the mean of the second half of trials. The DS examination does not address this analysis.

[2] Specifically, the standard deviation of participants' running means at trial *n* will be approximately $\frac{s}{\sqrt{n-1}}$.

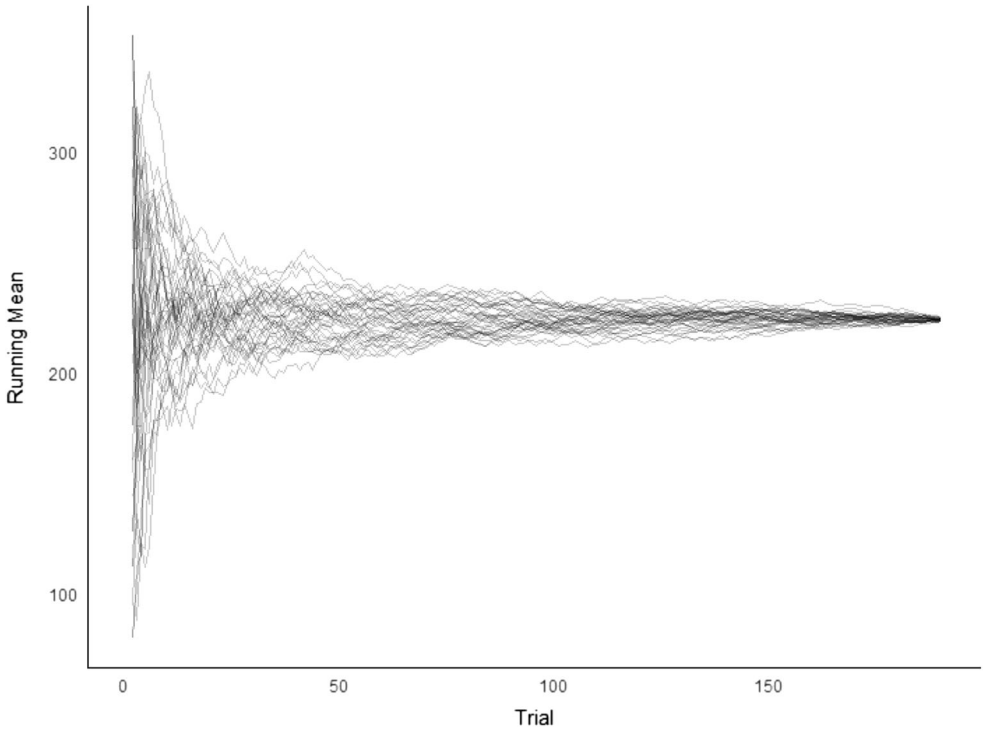

Fig. 2 Running mean for each participant in a simulated data set from Trials 2–190

the running mean, the DS approach is unable to detect its effect, although it does detect the effect of recent stimuli. In contrast, the deviations approach estimates both effects with comparable precision and does so in the second half of trials as well as it does in the first.

In response to concerns about power in their analysis, DS simulated data based on the changing distribution used in DHHC Experiment 1. In this simulation, each response was a weighted average of the target stimulus (.9) and the running mean (.1), with some added noise. For this data, the DS

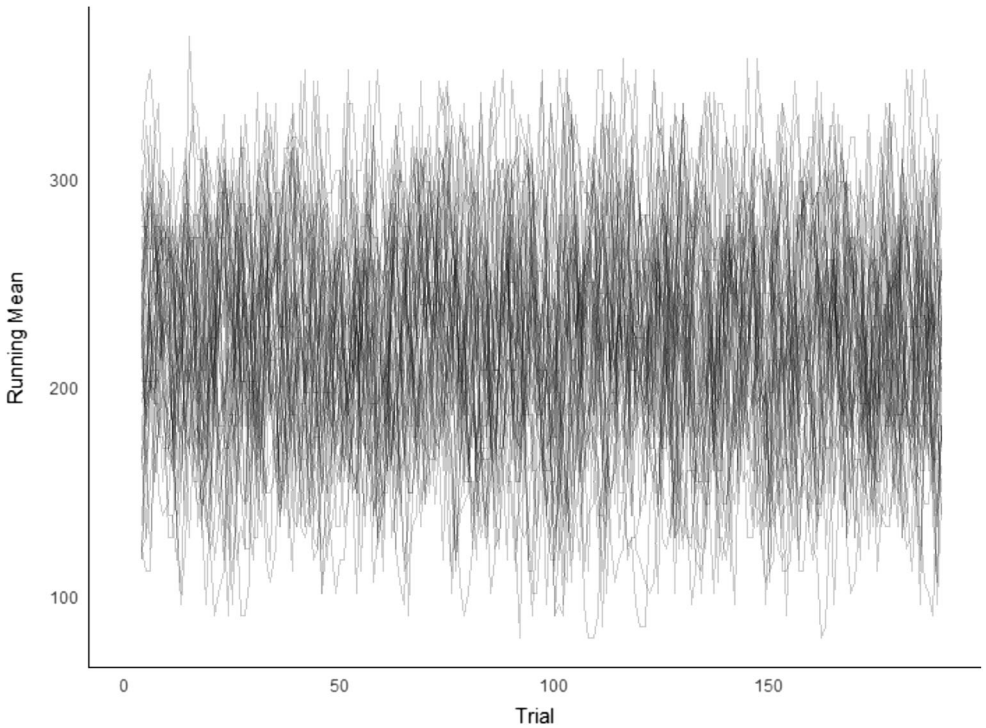

Fig. 3 Mean of previous three stimuli for each participant in a simulated data set from Trials 4–190

Table 1 Coefficients with standard errors for each half of trials in a simulated data set

DS analyses (df = 3517)	First half of trials	Second half of trials
Intercept	−2.162 (5.839)	12.252 (29.640)
Stimulus	0.801 (.005)***	0.799 (0.005)***
Running mean	**0.106 (.027)*** **	**0.048 (0.132)*ns***
Mean of last 3	0.100 (.009)***	0.099 (0.008)***
Deviations analysis (df = 3518)		
Intercept	−0.648 (.414)	0.075 (0.483)
Deviation from running mean	**−0.099 (.010)*** **	**−0.102 (0.010)*** **
Deviation from mean of last 3	−0.100 (.009)***	−0.098 (0.008)***

Note. Results of random-intercepts model fit with lme4 (Bates, Maechler, Bolker, & Walker, 2015) packages in R. ***$p < .001$

Table 2 Coefficients with standard errors for DHHC Experiment 1 and DHHC Experiment 2

DS analyses	DHHC 1	DHHC 2
df	4668	8457
Intercept	12.836 (5.875)*	19.815 (7.306)
Stimulus	0.803 (.005)***	0.784 (0.005)***
Running mean	**0.041 (.028)*ns***	**0.060 (0.032) *ns***
Mean of last 3	0.092 (.009)***	0.068 (0.009)***
Deviations analysis		
df	4670	8458
Intercept	−1.676 (2.262)	0.127 (2.582)
Deviation from running mean	**−0.113 (.009)*** **	**−0.147 (0.010)*** **
Deviation from mean of last 3	−0.085 (.008)***	−0.067 (0.009)***

Note. Results of random-intercepts model fit with lme4 package in R. ***$p < .001$, *$p < .05$

analysis produced a significant effect of the running mean on estimates, a result that they claimed, "should leave no doubt that our methods are able to detect CAM by identifying a significant relationship involving the running mean variable." However, as the right column of Table 1 illustrates, it is possible to produce simulated data with baked-in bias toward the running mean in which the DS analysis fails to identify the running mean effect. When both analytical approaches are applied to the data that DS generated, the deviations analyses estimates the simulated effect more precisely than does the DS analysis (see OSF page for analysis). This is not surprising, because the deviations analysis follows directly from the weighted average at the heart of CAM.

Applied to the real data from DHHC, the two analytical approaches produce different results. In each experiment, the DS analysis finds a significant effect of the last three stimuli,[3] but not of the running mean, whereas the deviation analysis finds an effect of both, with the effect of the running mean being larger than the effect of the recent stimuli (see Table 2). The deviations analyses also produces intercepts near zero, indicating that estimates are biased toward the hypothesized values (running mean and mean of last three) rather than toward some other value. The estimates here are very close to those presented in the original DHHC paper, which did not center the data but instead ran a regression through the origin (intercept set to zero) to estimate the relevant slopes.

Based on the null effect of the running mean in their analysis, Duffy and Smith concluded that DHHC was wrong and that there is no evidence of bias toward the running mean. An alternative interpretation is that this null effect shows the limitations of an analytical strategy. The deviations analysis provides more precise estimates of the running mean effect in simulated data, and when applied to real data, it finds a significant effect. Its results are compatible with Fig. 1, which suggests a process in which estimates are biased toward a long-run mean.

It is important to note that both analyses show a significant effect of the mean of the last three stimuli. DHHC failed to detect this effect, possibly because, as DS points out, they did not use a hierarchical regression approach and instead estimated parameters for individual participants. DHHC acknowledged that there may be an effect of recent stimuli and addressed situations in which giving more weight to recent stimuli would be sensible. Their conclusion was that a process of blending the current stimulus with only recent stimuli would not produce the pattern of bias observed in this data set.

Implications for the category adjustment model

As are all models, CAM is obviously wrong (Box, 1979). It makes assumptions that are simplifying and skeletal (e.g., treating trace memory and priors as normally distributed) and never claims to capture the whole complexity of human judgment. The DS paper constructs a straw man when it claims "CAM predicts that judgments will not be affected by features of the experiment that do not improve the accuracy of the judgment" (p. 1740). Their claim that "one prediction of CAM is that participants will not be sensitive to recently viewed stimuli" (p. 1740) should be revised to state that CAM makes no strong predictions about the effect of recent stimuli. DHHC concluded that recent stimuli were not driving the observed effects, but also acknowledged that in a systematically changing world, it could be adaptive to use a prior based on recent stimuli rather than one learned long ago (cf. Duffy & Crawford, 2008). More broadly, CAM does not require purely inductive priors in which every previous instance

[3] Here, I present an analysis using the mean of the last three stimuli. Many other subsets could be used to instantiate the idea of recency (as in DS) and the last three deserve no special status; they just happen to produce stronger effects than other subsets I examined.

is weighted equally, and it does not require that participants start every experiment with a diffuse, uninformative prior, as DS assume. What it (and Bayes theorem) does offer is a way of quantifying why such open-mindedness would be a good idea.

CAM has been useful as a descriptive rather than normative Bayesian model, a distinction suggested by Tauber, Navarro, Perfors, and Steyvers (2017). It has offered an explanation for certain biases in memory and generated novel predictions about how manipulations that affect the precision of a trace memory or prior would influence bias. Many studies have used CAM to examine the nature of the priors people actually use without requiring that those priors be optimally tuned to the environment. In Huttenlocher et al. (2000) and DHHC, the priors were thought to be the mean of all presented stimuli. In Duffy and Crawford (2008), the prior appeared to give more weight to stimuli presented early in the sequence. In many spatial memory tasks (e.g., Crawford, Landy, & Salthouse, 2016; Huttenlocher, Hedges, & Duncan, 1991), the priors people use have nothing to do with the stimulus distribution at all, and yet those priors still have a stronger influence under conditions that are likely to make trace memory less precise, as predicted by CAM.

Conclusion

I disagree with Duffy and Smith's (2018) conclusion that responses reported in the DHHC paper are biased toward only recent stimuli and not toward the running mean. I raise concerns about their analytical strategy, suggest an alternative approach, and show that this alternative better recovers the impact of the running mean in simulated data. Applied to the real data, this analysis suggests that estimates are biased toward the running mean, and to a lesser extent, toward recent stimuli. Finally, I suggest that these results, and the various null results presented by DS, do not have major theoretical significance for CAM and similar Bayesian models.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi:https://doi.org/10.18637/jss.v067.i01

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), Robustness in statistics (pp. 201–236). New York, NY: Academic Press.

Choplin, J. M., & Hummel, J. E. (2002). Magnitude comparisons distort mental representations of magnitude. *Journal of Experimental Psychology: General, 131*, 270–286. doi:https://doi.org/10.1037/0096-3445.131.2.270

Cornell, J. A. (2011). A primer on experiments with mixtures. New York, NY: Wiley.

Crawford, L. E., Landy, D., & Salthouse, T. (2016). Spatial working memory capacity predicts bias in estimates of location. *Journal of Experimental Psychology: Learning, Memory & Cognition 42*, 1434–1447. doi:https://doi.org/10.1037/xlm0000228

Duffy, S., & Crawford, L. E. (2008). Primacy or recency effects in forming inductive categories. *Memory & Cognition*, *36*(3), 567–577. doi:https://doi.org/10.3758/MC.36.3.567

Duffy, S., Huttenlocher, J., & Crawford, L. E. (2006). Children use categories to maximize accuracy in estimation. *Developmental Science, 9,* 597–603. doi:https://doi.org/10.1111/j.1467-7687.2006.00538.x

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*(2), 224–230. doi:https://doi.org/10.3758/PBR.17.2.224

Duffy, S., & Smith, J. (2018). Category effects on stimulus estimation: Shifting and skewed frequency distributions—A reexamination. *Psychonomic Bulletin & Review*, *25*(5), 1740–1750. doi:https://doi.org/10.3758/s13423-017-1392-7

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98,* 352–376. doi:https://doi.org/10.1037/0033-295X.98.3.352

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129,* 220–241. doi:https://doi.org/10.1037/0096-3445.129.2.220

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. doi:https://doi.org/10.1037/rev0000052

Weisberg, S. (1985). Applied linear regression. New York, NY: John Wiley & Sons.

Wickham, H. (2009). ggplot2: Elegant graphics for data analysis. New York, NY: Springer-Verlag.