Attention, Perception, & Psychophysics (2020) 82:1710–1726
https://doi.org/10.3758/s13414-019-01932-z

Context effects on reproduced magnitudes from short-term and long-term memory

Douglas H. Wedell [1] · William M. Hayes [1] · Jongwan Kim [2]

Published online: 2 January 2020

Abstract
Extant research has demonstrated strong contextual dependencies in reproducing magnitudes of perceptual stimuli from short-term memory. Two experiments examined how context as defined by (a) the mean of the distribution, (b) stimulus ranks, (c) values of anchor stimuli used in the reproduction task, and (d) values from the most recent trial operate on estimates of square size. Experiment 1 demonstrated distributional contrast effects on ratings of squares and distributional assimilation effects on reproduction of squares from short-term memory for the same participants. The fit of a modified version of the category adjustment model demonstrated reliable effects of the running mean, start anchors, and previous stimulus on reproduction bias. In Experiment 2, participants first learned to associate labels with squares, then reproduced square sizes based on the label cues, a long-term memory task, followed by a reproduction from short-term memory task as in Experiment 1. Results for the short-term memory task were largely consistent with Experiment 1. Results for the long-term memory task showed a very different pattern of effects, with larger reproduced sizes when squares were drawn from positively skewed rather than negatively skewed distributions. This contrast effect was explained by a modified range-frequency model as the result of rank encoding of square values along with displacement away from the running mean and shifts towards the prior response and start anchors. The combined results identify multiple sources of context effects in estimation that depend critically on memory retrieval factors and show how they can be incorporated into existing models.

Keywords Category adjustment model · Range-frequency theory · Visual memory · Bayesian modeling · Memory: Visual working and short-term memory · Memory: Long-term memory

Introduction

Estimation tasks involve describing objects or events on objective scales, such as when assessing age in years, temporal durations in seconds, heights in inches, distances in feet, or likelihoods in relative frequencies. Often estimates are made on a numerical scale, which can introduce biases due to how numbers are perceived and used (Poulton, Simmonds, & Warren, 1968). An alternative approach is to use a reproduction task in making estimates, so that reproduced properties can be directly compared to the stimulus properties. For example, one may be asked to reproduce a temporal duration recently experienced (Ryan, 2011) or the size of an object just viewed (Choplin & Hummel, 2002; Duffy, Huttenlocher, Hedges, & Crawford, 2010), or the location on the screen where a dot was recently viewed (Fitting, Wedell, & Allen, 2007; Huttenlocher, Hedges, & Duncan, 1991). These tasks often reveal systematic biases in reproduction that are used to make inferences about memory encoding and retrieval.

One of the long-standing contextual biases found in estimation tasks such as these is the central tendency effect. Hollingworth (1910) described how estimated values in both reproduction and perceptual matching tasks shift toward the mean of the contextual distribution presented to participants. This assimilative shift toward the mean has been found for numerous perceptual dimensions across multiple estimation methods (Helson, 1964; Helström, 1985; Jones & McAuley, 2005; Marks, 1992; Ryan, 2011). Why should estimates regress toward the central tendency of the contextual distribution of stimuli being judged? Two types of general accounts implicate either sequential processes (Choplin & Hummel, 2002, Duffy & Smith, 2018; Sailor & Antoine, 2005; Ward, 1979) or distributional processes (Duffy et al., 2010; Helson,

✉ Douglas H. Wedell
wedell@mailbox.sc.edu

[1] Department of Psychology, University of South Carolina, 1512 Pendleton Street, Columbia, SC 29208, USA

[2] Department of Psychology, University of Maryland, College Park, MD, USA

1964; Huttenlocher et al., 1991). According to sequential accounts, biases are based on concomitant changes in sequences of stimuli that result from the distributional changes. For example, a positively skewed distribution will result in a preponderance of low magnitude stimuli preceding a given target stimulus, but a negatively skewed distribution will result in the opposite relationship. If estimates shift toward the most recently experienced stimuli (as in first-order assimilation effects), then when estimates are averaged across trials they will consequently shift toward the central tendency of the distribution. On the other hand, distributional accounts argue that the central tendency of the distribution is represented directly in memory, updated on each trial, and used in judgment and estimation (Helson, 1964). Accordingly, distribution effects should be found even after sequential effects are accounted for.

One distributional model that has been successfully applied to reproduction of stimulus magnitudes and spatial locations is the Category Adjustment (CA) model developed by Huttenlocher and colleagues (Crawford, Huttenlocher, & Engebretson, 2000; Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Hedges, & Vevea, 2000). The CA model has enhanced our understanding of the central tendency effect in several ways and has been successfully applied to memory for both unidimensional and multidimensional stimuli. The model conceives of bias effects within a Bayesian framework, with shifts of estimates toward a category prototype posited to be adaptive in that they produce less overall error in estimation. This is because when there is uncertainty about the exact value of a target stimulus, biasing one's estimates in the direction of the central tendency of the contextual distribution (the "prior") will minimize average estimation error (Duffy et al., 2010).

Much of the research studying categorical and distributional effects has been conducted using a paradigm in which the stimulus to be reproduced is briefly presented followed by a few seconds of a mask or blank screen before one engages in reproduction or estimation (see Fitting, Wedell, & Allen, 2008, for a study manipulating presentation and delay times). In this experimental paradigm the stimulus is presumably maintained in short-term visual memory and so its physical characteristics should be fairly accurately remembered. The CA model proposes that uncertainty about the stimulus value is resolved by adjustment toward the category prototype.

What happens when the stimulus representation must be retrieved from long-term memory? Choplin and Wedell (2014) explored this question by having participants memorize calorie information for seven hamburgers with unique descriptions drawn from a positively or negatively skewed distribution. After a 1-min distractor task, participants recalled the number of calories for each burger in a numerical reproduction task. Across three replications, the pattern consistently reflected strong contrast effects for middle values and assimilation effects for extreme values. Contrast effects are defined as shifts of judgments or estimates away from contextual values and assimilation effects as shifts of judgments or estimates toward contextual values (Wedell, Hicklin, & Smarandescu, 2007). This overall pattern of assimilation for extreme stimuli and contrast for moderate stimuli was explained by two different models, the comparison induced distortion (CID) model (Choplin, 2007; Choplin & Hummel, 2002) and a modification of range-frequency theory (RFT) (Parducci, 1995). According to the CID model, which is typically applied sequentially, the judge compares the difference in a pair of stimulus values to an intermediate interval and distorts small differences to be larger and large differences to be smaller. According to the modified range-frequency (RF) model, stimulus values are remembered by their ranks and then these rank values are mapped onto the appropriate stimulus range in order to reproduce the stimulus value. For a skewed stimulus distribution, this process produces contrast effects when the judge estimates the magnitude of middle stimuli. Because the standard RFT does not explain assimilation effects at the extremes, it was modified to include a weighting of the distributional mean as in the CA model.

Note that when endpoints are matched across contexts and skewing is manipulated, RFT explains contrast as resulting from the differences in stimulus ranks rather than differences in distribution means. In a skewed distribution, the magnitudes of stimuli in the dense region are closer together than the magnitudes of stimuli in the tail. A rank transformation of magnitudes will then overestimate differences in dense regions and underestimate differences in sparse regions (Wedell, 1996). Consequently, the same stimulus magnitude will be judged greater in the positively skewed distribution than in the negatively skewed distribution, with the contrast effect greatest for middle values. In a reproduction task, if individuals remember ranks instead of stimulus magnitudes directly, RFT predicts that they will overestimate smaller magnitude differences and underestimate larger magnitude differences. The end result is that the same moderate stimulus will be reproduced larger when embedded in a positively skewed distribution than when embedded in a negatively skewed distribution, i.e. a contrast effect in reproduction (Choplin & Wedell, 2014).

The extant research then points to several contextual factors possibly influencing estimation of values in a reproduction task. According to the CA model, the category or distributional mean should result in an assimilative bias (Crawford, 2019). An alternative explanation of these distributional effects is based on sequential effects (Duffy & Smith, 2018; Sailor & Antoine, 2005). Studies involving reproduction of magnitudes such as line length or square size typically proceed from a starting or an anchor stimulus (large, moderate, or small) that is adjusted. There is a large literature on assimilative effects created by anchoring and insufficient adjustment

that implicates these start values as another contextual influence on the reproduced value (for reviews of anchoring effects see Epley & Gilovich, 2006; Mussweiler & Strack, 1999). Finally, the direction of the effect (assimilation or contrast) may depend on whether relevant stimulus magnitudes are readily available or not. When the stimulus has been recently presented, fine-grain magnitude information may be readily retrieved from short-term memory and assimilation effects may predominate. However, when magnitude information must be stored and retrieved from long-term memory, fine-grain magnitude information may be more fragile and difficult to recover, whereas gist information concerning relative magnitudes may be more robust and readily accessible (Reyna & Brainerd, 1995). Rank encoding represents a type of gist that may be retrieved from long-term memory and used in evaluation and estimation processes that produce contrast effects (Choplin & Hummel, 2002; Stewart, Chater, & Brown, 2006; Ungemach, Stewart, & Reimers, 2011; Wedell, 1996).

The two experiments reported here investigate these sources of contextual bias in a square size reproduction task. Experiment 1 used a standard short-term memory reproduction task and tested for distributional, sequential, and anchor effects. Participants also rated magnitudes on a subjective scale to verify that rank-based encoding of stimulus magnitudes produces contrast in line with RFT, whereas reproduced estimates of magnitude from short-term memory produce assimilation in the same participants. Experiment 2 included a learning phase in which participants learned to identify different square sizes associated with different letters and then reproduced square sizes from letter cues, a task based on retrieval from long-term memory. For purposes of direct comparison of the two memory retrieval conditions, participants also reproduced the squares using the standard short-term memory paradigm. The overarching goal was to develop a clearer understanding of the sources of bias and the direction of bias in estimation tasks performed under different memory constraints designed to make fine-grain magnitudes and stimulus rank information differentially accessible.

Experiment 1

The aim of Experiment 1 was to examine how distributional effects (bias toward the running mean), anchor effects, and sequential effects operate when reproducing simple perceptual stimuli from short-term memory. Our paradigm provided a strong test of anchor effects by varying the start-anchor stimulus within participants and across different blocks of trials. Additionally, we fit a computational model that accounts for the repeated-measures nature of our data in order to provide a more sensitive test for the presence of sequential effects (e.g., Crawford, 2019; Duffy & Smith, 2018). The CA model predicts that a main source of bias should be the running mean of the stimulus distribution since it would minimize the mean square error of the estimates (Duffy et al., 2010). The demonstration of large anchor effects (Cornsweet, 1962; Epley & Gilovich, 2006; Mussweiler & Strack, 1999; Tversky & Kahneman, 1974) and sequential effects (Duffy & Smith, 2018; Sailor & Antoine, 2005) in this experiment would thus be inconsistent with a strictly Bayesian interpretation of reproduction bias, and would suggest that the CA model should be at least modified to incorporate other sources of bias. One way to do so is to assume the prototype is a weighted average of the running mean, recent values, and anchors, even though the anchors are not representative of the distribution.

The present experiment also included category ratings of square size in a set of trials either before or after the estimation set. Whereas estimations assimilate toward the central tendency of the distribution, category ratings tend to show an opposite, contrastive shift in which rated values reflect stimulus ranks within the contextual distribution (Wedell, 1996). Including category ratings allowed us to examine how the presence of assimilation or contrast effects depends on the nature of the task, objective reproduction, or subjective valuation (Biernat & Manis, 2007). These rank-based ratings are assumed to represent the gist encoding of the stimuli that could be used to reconstruct stimulus magnitudes when fine-grain information is difficult to retrieve, as tested in Experiment 2.

Method

Participants and design

Participants were 48 undergraduates (35 women, 13 men) from the University of South Carolina who received course credit for their voluntary participation. They were randomly assigned to one of eight conditions resulting from the between-subjects factorial combination of distribution (positive or negative skew), task order (ratings first or estimations first), and anchor order (small followed by large anchor in alternate blocks, or the reverse order). Within-participant variables included task (rating and estimation), anchor stimulus (large and small), and stimuli (seven squares sizes common to both distributions).

Materials and apparatus

All instructions and stimulus materials were presented via desktop computers (13 in. screens in a 640 × 480 pixel array) and all responses were recorded on the computers. Squares were solid blue on a white background projected at the center of the screen and varied in width in increments of 20 pixels (20, 40, 60, 80, 100, 120, and 140). The small anchor was 10 pixels wide and the large anchor was 150 pixels wide, each

just outside the range of the experimental series. In the positively skewed distribution, the frequency of the seven square sizes in each block of 11 trials was 2, 3, 2, 1, 1, 1, and 1, respectively. In the negatively skewed distribution, the frequencies were reversed, 1, 1, 1, 1, 2, 3, and 2, respectively. Participants used the arrow keys, number keys, and spacebar to make responses and were tested in groups of up to six in a large room, with computers spaced approximately 1 m apart.

Procedure

Instructions stated that the experiment was about how people make ratings and estimations of square size. Depending on task order, participants were then given instructions for either rating or estimation. Instructions for rating presented a 9-point scale, with end points labeled "very small" and "very large," respectively, and stated that each square would appear briefly on the screen, followed by the rating scale. Participants were to input the number between 1 and 9 that indicated how large they felt the square appeared to them. Instructions for estimation told participants that each square would appear briefly on the screen, followed by another square that they were to adjust in size until they felt it matched the size of the previous square. They used the up and down arrow keys to adjust the size and then pressed the space bar to record their estimation. They were told that after every 11 trials they would be given feedback about how accurate they were in that block of trials.

For both tasks, a trial began with the prompt for the participant to press the space bar, after which the square appeared for 1 s followed by a 0.5-s blank screen. For rating trials, the rating scale appeared and participants pressed one of the number keys to record their response. The response was echoed on the screen for 0.5 s and then the screen was blank for 0.5 s before a new trial began. For estimation trials, the anchor square appeared after the blank interval along with instructions on the screen reminding participants to use the arrow keys to adjust the square size larger or smaller until they were satisfied it matched the previous square and then hit the space bar to record the estimate. Square size changed by 1 pixel for each press of the arrow key, and holding a key down caused the square to change size rapidly. After each block of 11 trials, the accuracy percentage over the last 11 trials was reported to the participant. An estimate was counted accurate if it was within 5 pixels of the actual width. After each accuracy summary, the anchor square was changed for the next block of 11 trials. In total, there were six blocks of 11 trials for each participant in both the estimation and the rating tasks.

Data analysis and modeling

Bayesian ANOVA To analyze the estimation and rating data at the group level, we employed Bayesian repeated-measures ANOVAs with random effects for each participant. Bayesian ANOVAs have many advantages over conventional frequentist ANOVAs, including the ability to provide evidence in favor of null hypotheses and the ability to state evidence for different hypotheses in a graduated fashion without having to make binary decisions, such as rejecting or failing to reject a hypothesis (Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017). We used the default prior specification for factorial designs implemented in Morey and Rouder's (2014) R package BayesFactor (for a tutorial, see Rouder et al., 2017). The number of Monte Carlo simulations was set to 50,000 for numerical approximation of the Bayes factors.

Bayes factors provide a natural way of comparing two competing hypotheses or models based on the amount of evidence for each in the data. More specifically, if M_1 is a model in which an effect is present and M_0 is a model in which the effect is constrained to be absent, the Bayes factor in favor of M_1 is given by

$$\mathrm{BF}_{10} = \frac{\Pr(D|M_1)}{\Pr(D|M_0)}. \quad (1)$$

In Eq. 1, $\Pr(D|M)$ denotes the marginal probability of the data D conditional on model M, averaged over all possible parameter combinations. For example, a Bayes factor of 20 would indicate that the data were 20 times more likely under M_1 than under M_0. A Bayes factor between 1 and 3 is considered "weak" evidence in favor of M_1, between 3 and 10 "moderate" evidence, between 10 and 30 "strong" evidence, between 30 and 100 "very strong" evidence, and greater than 100 "extreme" evidence for M_1 (Wagenmakers et al., 2018).

In addition to reporting Bayes factors, we also report 95% Bayesian credible intervals (BCIs) for means, mean differences, and polynomial contrasts of interest. The bounds of the 95% BCI are the quantiles that contain the middle 95% percent of the posterior distribution, meaning there is a 95% probability that the parameter of interest falls within those bounds, given the model (Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2013).

Computational modeling We used a hierarchical Bayesian model in order to simultaneously describe effects at both the aggregate and the individual level and to characterize the underlying cognitive processes that might have produced the observed data (Lee, 2011; Shiffrin, Lee, Kim, & Wagenmakers, 2008). Our approach assumes that each individual's parameters are drawn from group-level distributions, and the parameters governing these group-level distributions are given independent prior distributions. Thus, in using the data to update prior beliefs, the hierarchical Bayesian approach yields posterior distributions for both individual-level and population-level parameters. Our primary focus was on the population-level parameters.

We obtained samples from the joint posterior of the model parameters with Stan (version 2.19), a programming language that uses Hamiltonian Monte Carlo (HMC), an extension of the Markov Chain Monte Carlo algorithm (MCMC), to sample from complex probability distributions (Carpenter, Gelman, Hoffman, & Lee, 2016). Models were coded as Stan programs and all sampling was carried out in RStan (Stan Development Team, 2016). Individual-level parameters were assumed to be drawn from normal group-level distributions, and we used a non-centered parameterization to aid sampling efficiency (Betancourt & Girolami, 2015) (see *Results* for an example of this parameterization).

Results

Reproduction bias

In line with the typical analyses of these types of data, the actual stimulus value was subtracted from the response to create a bias score, with positive values reflecting overestimates of size and negative values reflecting underestimates of size.[1] In the standard CA model analysis of bias functions, the point at which the downward sloping function intersects the 0-bias value represents the category prototype. Figure 1 presents mean reproduction biases segregated by distribution and anchor (note that both axes were scaled down by a factor of 10 to agree with our modeling analysis, see below). For both small and large anchors, the bias function for the positively skewed distribution is shifted to the left of that for the negatively skewed distribution, reflecting the predicted assimilation toward the distribution mean. However, a much larger bias effect is found for the start anchors themselves. For both the positively and negatively skewed distributions, the bias function for the small anchor is displaced to the left of the bias function for the large anchor, a strong assimilative effect toward the anchor stimulus. Note also that while the bias functions are negatively sloped, they also follow a clear cubic trend that suggests a diminishing of bias effects near end stimuli. We first verified these effects at the group level using ANOVA and then attempted to explain them using computational modeling.

Bayesian ANOVA To evaluate these effects, the bias scores were submitted to a 2 (Skew) × 2 (Anchor) × 7 (Stimulus) Bayesian repeated-measures ANOVA. There was extreme evidence in favor of an Anchor main effect ($BF_{10} = 5.11 \times 10^{35}$). When the anchor stimulus was large, reproduced squares were larger than when the anchor stimulus was small (95% BCI for the difference [6.95, 9.22]). There was also extreme evidence for a Stimulus main effect ($BF_{10} = 7.56 \times 10^{30}$) and strong evidence for an Anchor × Stimulus interaction ($BF_{10} = 15.02$). An analysis of posterior samples of the means revealed a more negative quadratic trend across square sizes when the anchor was large compared to when it was small (95% BCI for the difference [-8.31, -2.41]). There were strong linear and cubic trends for both anchors, reflecting the overall tendency to underestimate large squares and overestimate small squares, as well as the reduced bias near end stimuli. The evidence in favor of a Skew effect was weak but positive ($BF_{10} = 2.17$). On average, reproduced squares were larger when the distribution of squares was negatively skewed (95% BCI of the difference [0.29, 5.00]), consistent with an assimilative shift toward the central tendency of the distribution. However, the presence of strong anchor effects suggests that participants also recruited irrelevant contextual information when estimating the squares.

Computational modeling We used computational modeling to examine the potential sources of bias that might have contributed to the Skew effect in the previous analysis. The two potential sources of the Skew effect that we considered were the running mean of the distribution and the stimulus presented on the previous trial. We also examined the anchor effect on a trial-by-trial basis and included a mechanism for explaining the diminishing bias near end stimuli.

The model was fit to the aggregated trial-by-trial data after excluding the first trial for each participant. The structure of the model built on the basic CA model described by Huttenlocher and colleagues (Duffy et al., 2010; Huttenlocher et al., 1991):

$$B_{ijt} = \lambda_{ij} S_j + \left(1-\lambda_{ij}\right)\left(w_{RM,i} RM_t + w_{AS,i} AS_t + w_{PREV,i} S_{t-1}\right) - S_j. \quad (2)$$

In Eq. 2, B_{ijt} is the bias in individual i's estimate of the j^{th} stimulus on trial t. The fine-grain weighting parameter λ_{ij} ($0 \leq \lambda_{ij} \leq 1$) is the relative weight that individual i gives to the actual stimulus value S_j, while $1 - \lambda_{ij}$ is the weight that individual i gives to the category prototype. The model assumes that the category prototype is comprised of three potential sources of bias: the running mean of the contextual distribution experienced up to the t^{th} trial (RM_t), the size of the starting anchor on the t^{th} trial (AS_t), and the size of the stimulus on the previous trial (S_{t-1}). Each of these three components is weighted by an individual-level coefficient to allow for variability across participants in the strength of each source of bias.

Figure 1 shows a clear cubic trend in which bias diminishes for stimuli near the end points of the distribution. Haun, Allen, and Wedell (2005) observed a similar cubic component to bias

[1] Outlier trials, defined as those in which reproduction bias was more than three standard deviations higher or lower than the group average (1.3% of all trials), were excluded from the following analyses.

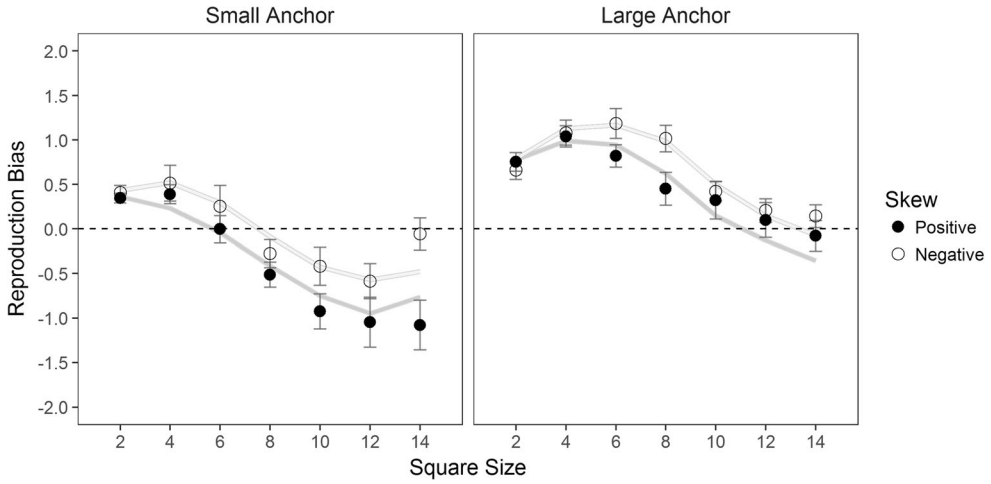

Fig. 1 Mean reproduction bias as a function of skew, anchor, and stimulus size (Experiment 1). Lines show the fit of the hierarchical Bayesian modified CA model (Eq. 2). On each trial, the model prediction is the expectation of the posterior predictive distribution, given the covariates for that trial. Both axes were scaled down by a factor of 10 to match the modeling analysis. Error bars represent one standard error of the mean

and attributed this to the bowing effect in perception in which stimulus discrimination is better near end stimuli. The CA model posits that weighting of the prototype should decrease when memory for fine-grain information (i.e., stimulus discriminability) increases. Hence, the fine-grain weighting parameter was assumed to follow a quadratic function so that λ is higher near the end stimuli:

$$\lambda_{ij} = a_i + b_i\left(S_j - 8.0\right)^2. \tag{3}$$

As seen in Eq. 3, weighting of fine-grain information varies trial to trial based on the squared distance of the j^{th} stimulus from the (scaled) midpoint of the overall distribution (8.0).[2] The a_i parameter represents individual i's weighting of fine-grain information when the stimulus is at the center of the distribution, while the b_i parameter modulates the weighting of fine-grain information for stimuli closer to the endpoints of the distribution.

The individual-level parameters in this model were assumed to be drawn from group-level normal distributions. Means and standard deviations for the group-level distributions were given vague normal and half-normal priors, respectively. For example, the individual-level weights for the running mean were modeled as follows:

$$\begin{aligned} &\mu_{w_{RM}} \sim \text{Normal}(0, 20) \\ &\sigma_{w_{RM}} \sim \text{Half}-\text{Normal}(20) \\ &z_{w_{RM,i}} \sim \text{Normal}(0, 1) \quad (i = 1, \ldots, N) \\ &w_{RM,i} = \mu_{w_{RM}} + \sigma_{w_{RM}} \cdot z_{w_{RM,i}} \quad (i = 1, \ldots, N) \end{aligned} \tag{4}$$

We ran four chains with 2,500 samples each and discarded the first 1,250 as warm-up samples, resulting in a total of 5,000 posterior samples for each parameter. The $\hat{R}$ statistics for each parameter, which represent a ratio of between- to within-chain variability, were all below 1.01, suggesting that the chains had each converged to the posterior distribution (Gelman et al., 2013).

The hierarchical model provided an adequate fit to the group-averaged data, with slight misfit to the mean biases for the largest squares (Fig. 1).[3] We examined the posteriors for the group means of the weights of the three sources of bias in the category prototype (w_{RM}, w_{AS}, w_{PREV}) and found that they were all well above zero (Fig. 2). That is, the previous stimulus (95% BCI [0.40, 0.62]), anchor stimulus (95% BCI [0.32, 0.56]), and the running mean of the distribution (95% BCI [0.12, 0.52]) were each positively and robustly associated with reproduction bias. Thus, while there is convincing evidence that participants' estimates were biased toward the mean, there was also very strong evidence for a bias toward irrelevant anchors and the most recently experienced stimulus.[4] The posterior means for the fine-grain weighting parameter λ were largest for the smallest and largest squares in the set (Fig. 2), consistent with discrimination being greatest near the end stimuli and, conversely, bias being greatest for squares closer to the midpoint of the distribution. This explanation can account for the cubic trend observed in Fig. 1 (Haun et al., 2005).

[2] To aid sampling efficiency, bias scores and all inputs were scaled down by a factor of 10.

[3] Code for fitting the model as well as plotting fits to individual participants' data can be found at https://osf.io/8abj4/.

[4] Given the static nature of the stimulus distributions in this experiment (as opposed to a shifting distribution, as in Duffy et al., 2010), it did not make sense to try and disentangle a running mean bias from a bias toward the mean of the most recent N stimuli (e.g., Duffy & Smith, 2018). For our distributions, the correlation between the mean of the previous N stimuli and the running mean becomes larger in magnitude as N increases, climbing above 0.90 when $N = 2$.

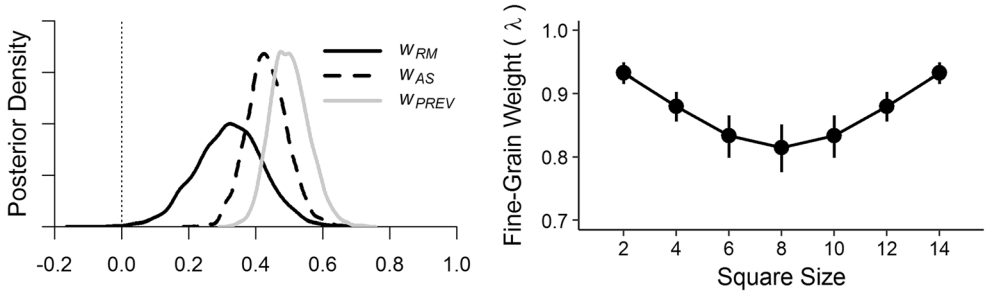

Fig. 2 **Left panel:** Posteriors for the group means of w_{RM}, w_{AS}, and w_{PREV}, i.e. the weights given to the running mean, anchor stimulus, and previous stimulus in the category prototype. **Right panel:** Posterior means for the fine-grain weighting parameter across different square sizes. Bars represent 95% Bayesian credible intervals

Ratings

Figure 3 presents the mean ratings of the square sizes segregated by skewing condition. The key result is that the large effects of context are in the opposite direction for ratings (contrast) than for estimation (assimilation). This finding was supported by a 2 (Skew) × 7 (Stimulus) Bayesian ANOVA with a random effect for each participant. There was extreme evidence for the Skew effect (BF_{10} = 1.12 × 10^{15}). Squares were rated as larger when they were drawn from the positively than from the negatively skewed distribution (95% BCI for the difference [0.45, 1.03]). There was also extreme evidence for the Stimulus main effect (BF_{10} = 1.56 × 10^{215}) and the Skew × Stimulus interaction (BF_{10} = 9.10 × 10^{10}). Although there was a strong linear trend across stimuli in both distributions, reflecting the clear tendency to assign higher ratings to larger squares, the trend was steeper for the positively than for the negatively skewed distribution (95% BCI for the difference [0.42, 0.92]). Consistent with RFT, the means for the positively skewed distribution had a negative quadratic trend (95% BCI [-0.62, -0.25]), while the means for the negatively skewed distribution had a positive quadratic trend (95% BCI [0.18, 0.53]). Finally, there was a more positive cubic trend for the negatively compared to the positively skewed distribution (95% BCI for the difference [0.04, 0.53]).

Figure 3 shows the fit of a modified RF model to the rating data (see Appendix for details). The model accounts for the opposing quadratic trends between the skew conditions by assuming that individuals track the percentile ranks of the stimuli as they encounter them and recruit these ranks when rating the size of each successive square. The weight of ranks, which varies between zero and one, was close to 0.50, as described in the Appendix.

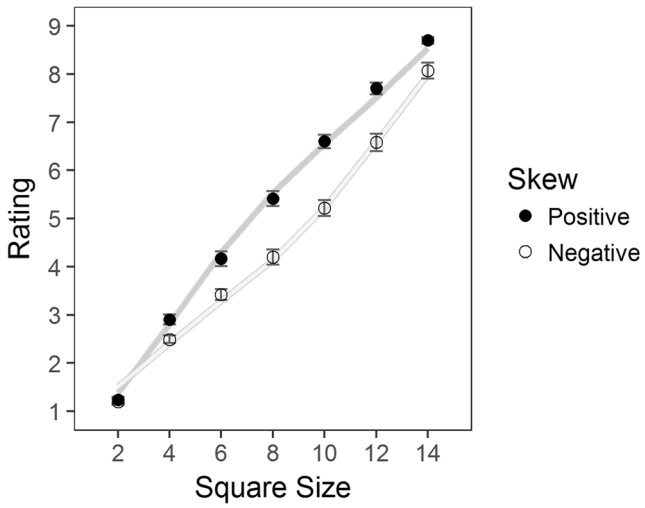

Fig. 3 Mean size ratings as a function of skew and stimulus size. Lines show the fit of a modified RF model (see Appendix). On each trial, the model prediction is the mean of the posterior predictive distribution, given the covariates for that trial. Square sizes were scaled down by a factor of 10 to match the modeling analysis. Error bars represent one standard error of the mean

Discussion

The CA framework was used for testing three sources of bias in stimulus reproduction from short-term memory by modeling the category prototype on a given trial as a function of the running mean, the size on the previous trial, and the size of the starting anchor stimulus. The CA model claim is that estimation functions in a way that is largely consistent with a Bayesian process in which the mean of the relevant category values is combined with information from the stimulus to provide a biased estimate that nevertheless reduces errors (Duffy et al., 2010). Other models have attributed reproduction biases to sequential dependencies (Choplin & Hummel, 2002; Duffy & Smith, 2018; Sailor & Antoine, 2005) or more general biasing processes, such as anchoring and adjustment (Tversky & Kahneman, 1974).

In partial support of the CA model, our results demonstrated systematic assimilation of reproduced stimuli toward the distribution mean. Consistent with some recent work in this area (Crawford, 2019), we found that this bias toward the running mean of the contextual distribution remained in effect even after accounting for effects of the previous stimulus. But it should be noted that with our between-subjects design and static distributions, it is difficult to disentangle the effects of

the running from the effects of the mean of the previous N stimuli, especially when $N \geq 2$. Although we found some evidence for a bias toward the running mean of the distribution, more research should be conducted with changing or shifting distributions to decouple the running mean from the mean of recent stimuli (e.g., Duffy et al., 2010).

Importantly, our results showed very strong shifts toward the values of anchor stimuli used as start values for making estimates. These results stand in contrast to those of Duffy et al. (2010), who reported no effect of the starting position. From a Bayesian standpoint, there is no reason to weight the anchor stimulus in the estimation process, unless it could be conceived as derived from the same population as the sampled stimulus. Some researchers have argued that anchoring and adjustment may provide a plausible psychological mechanism by which one can approximate averaging processes typically found in judgment (Wedell & Senter, 1997). One possibility then is that participants used anchoring and adjustment processes that included recent stimuli and start values as anchors. The inclusion of recent stimuli as anchors helps to achieve the normative (Bayesian) weighting of the relevant category mean and thus is broadly consistent with the CA model. The inclusion of the anchor stimuli used to start the adjustment process is not consistent with the Bayesian model, but it is not surprising given the large literature on such effects (Epley & Gilovich, 2006).

The present experiment further demonstrated that for these data, skew effects could be partially explained by participants recruiting the magnitude of the stimulus that appeared one trial back in the estimation process. These results are consistent with those reported in another square size estimation study (Sailor & Antoine, 2005), as well as recent re-examinations of the original Duffy et al. (2010) experiment (Crawford, 2019; Duffy & Smith, 2018). More research using shifting distributions is needed to clarify the conditions under which distributional effects on bias may be explicable in terms of sequential effects tied to recent stimuli.

Finally, it is worth noting that while estimations of size showed assimilative effects of context, evaluations of size in the form of category ratings showed contrastive effects of context. The demonstration of the different effects of context with these tasks suggests that neither task provides an unequivocal view of the stimulus representation, but rather that the representations may change flexibly with context and task. This finding is also consistent with the idea that estimating objective quantities (e.g., size) will often elicit assimilation effects while subjective ratings will often elicit contrast (Biernat & Manis, 1994; Biernat, Kobrynowicz, & Weber, 2003). Experiment 2 was designed to investigate a related issue, namely, whether stimulus representations might also change flexibly when the task involves cued retrieval from long-term memory instead of immediate retrieval from short-term memory.

Experiment 2

Experiment 1 was based on procedures typically used in reproduction tasks in which the participant reproduces each stimulus shortly after its presentation (Duffy et al., 2010; Huttenlocher, et al., 1991). Given the very brief time interval between presentation and recall typically used, these experiments examine biases linked to encoding into or retrieval from short-term memory. To what extent do the assimilation effects typically found for estimation from short-term memory transfer to estimation from long-term memory?

There are reasons to believe similar processes are applied when estimating values stored in long-term memory. First, in a spatial memory task using a human analog to the Morris water maze, Fitting, Allen, and Wedell (2007) had participants reconstruct one or three remembered locations after a few minutes delay, a week's delay, and 2 weeks' delay. For the critical target, they found assimilative effects on memory to cue-defined prototypes even after these very long delays. Thus, bias in reproduction from long-term memory may follow a similar assimilative shift toward category prototypes as found in the short-term memory version of a spatial memory task. Second, the typical finding within the stereotyping literature is that values of ambiguous stimuli tend to be biased toward stereotypic values or category means (Hilton & von Hippel, 1996, although for exceptions see Biernat & Manis, 2007). A result of these assimilative tendencies is that the similarity of category exemplars to one another is enhanced.

Alternatively, if one's goal is to remember specific values of exemplars linked to name cues, then encoding schemes that enhance the distinctiveness of exemplars may be used so as to reduce confusion at retrieval. To this end, rank-based encoding may enhance discrimination. Parducci (1995) has pointed out that the frequency principle in RFT works to maximize information transmitted about stimuli within a set, with enhancement of differences in denser regions of the distribution. These enhanced differences have been shown to result in decreased pairwise similarity (Wedell, 1996) as well as increased discriminability (Wedell, 2008). It is therefore reasonable to assume that when learning identities of different stimuli, rank information will be emphasized and encoded in memory. In support of this assertion, Pettibone and Wedell (2007) found that when participants had to learn identities of schematic faces linked to different groups (gnomes and leprechauns), category ratings of feature sizes reflected ranks within the groups and produced contrast effects. Ratings of pleasantness of faces showed an assimilation of the ideal point toward the mean of each category distribution. Hicklin and Wedell (2007) replicated and extended these findings by showing that the extent of contrast or assimilation on feature ratings depended on how well participants individuated exemplars within each group.

The results from these category-learning experiments (Hicklin & Wedell, 2007; Pettibone & Wedell, 2007) demonstrate that category ratings based on associated name cues can produce contrast on evaluations of features consistent with rank-based encoding. However, they do not speak to whether similar effects will be obtained for reproduction of stimulus values. As shown in Experiment 1, when no learning phase occurs and stimuli are retrieved from short-term memory, estimation from reproduction and magnitude ratings show opposing effects. However, the Choplin and Wedell (2014) experiment does speak directly to this issue. They found that the remembered calories for a given burger tended to shift away from the distributional mean as predicted by rank-based encoding. This occurred for moderate values, but end stimuli appeared to show an assimilative shift instead. Based on that study, it is reasonable to predict that analog reproduction of square sizes generated from learned label cues will result in a similar pattern of effects. If rank encoding captures the gist of the relative sizes within a distribution, then it may be more readily retrieved and used in estimation from long-term memory.

Experiment 2 provides for a direct comparison of the reproduction biases found for these two types of memory cues in a procedure that consisted of three phases. In Phase 1 participants learned to associate labels with each square size drawn from either a positively or a negatively skewed distribution. In Phase 2, they were presented with these labels and asked to recall the corresponding square size through reproducing the square on the screen, starting from either a large or a small anchor square. In Phase 3, they were presented each square again and reproduced the sizes after a 1-s delay, similar to Experiment 1. Our primary prediction is a dissociation of bias found for reproduction from short-term and long-term memory. We predict an assimilation to the mean for estimates from short-term memory, but a contrast effect reflecting rank-based encoding for estimates based on cued retrieval of sizes from long-term memory. A secondary prediction is that the assimilative effects of the previous trial and of the start anchor should be found for both reproduction tasks.

Method

Participants and design

Participants were 55 undergraduates (40 women, 15 men) from the University of South Carolina who received course credit for their voluntary participation. They were randomly assigned to one of four conditions resulting from the between-participants factorial combination of distribution (positive or negative skew) and anchor stimulus (large or small) (anchor stimulus was manipulated between participants due to the relatively small number of trials for the two reproduction tasks). Within-participant variables included square size (five squares sizes common to both distributions) and memory cue (square or label).

Materials and apparatus

All instructions and stimulus materials were presented via desktop computers (15-in. screens in a 640 × 480 pixel array) and all responses were recorded on the computers. Common squares to both distributions were varied in width in increments of 30 pixels (40, 70, 100, 130, and 160). The additional square sizes for the positively skewed distribution were 50, 60, 80, and 90 pixels and the additional sizes for the negatively skewed distribution were 110, 120, 140, and 150 pixels, so that each distribution had nine squares. The ordinal labels for squares were A, B, C, D, E, F, G, H, and I. The minimums and the maximums of both positive and negative distributions were equivalent (40 and 160 pixels, respectively) whereas the means were different (86.7 and 113.3 pixels, respectively). The small anchor was 30 pixels wide and the large anchor was 170 pixels wide, each just outside the range of the experimental series. Participants used the mouse buttons to make responses and were tested in groups of up to six in a large room, with computers spaced approximately 1 m apart.

Procedure

Instructions stated that the first task was to learn to associate letters with different square sizes, labeled A to I from smallest to largest. In a preview to the learning phase, each square was randomly presented twice along with its corresponding letter label, and participants were asked to input the letter. In the self-paced learning phase, the same squares were randomly presented in eight blocks without the accompanying label, and the participants were asked to indicate which square was presented by pressing the corresponding key (A–I). Feedback indicated a running accuracy rate and indicated the correct response if the participants gave an incorrect response.

The next phase was a name-based reproduction task that required retrieval of square sizes from long-term memory. This was the next task for all participants because it was the focus of our investigation and it facilitated memory for square-label pairings. After clicking the mouse to start a trial, a label cue indicating the target square along with an adjustable anchor square appeared. Two boxes labeled “Click Here to Make Smaller” and “Click Here to Make Larger” were presented and participants clicked on one box to make the square smaller and the other box to make it larger. Square size changed by one pixel for each click of the mouse, and holding a mouse button caused the square to change size rapidly. When they decided that the size matched the size of the square in memory, they could click the “Done” symbol on the screen to proceed

to the next trial. Square labels were randomly presented within each of three blocks of nine trials for a total 27 trials.

The last phase was an immediate reproduction task that required retrieval of square sizes from short-term memory. In this task, each square appeared for 1 s after the mouse click to begin a trial, followed by a 1-s blank screen. After the presentation, the anchor square appeared on the screen in an offset location and participants were required to adjust its size until it matched the square held in memory. Square sizes were randomly presented within each of three blocks of nine trials for a total 27 trials.

Results

Learning phase

Table 1 shows the proportion of times each label was assigned to each square in the initial learning phase. Participants in both skew conditions were most accurate in identifying the smallest and largest squares (A and I), with accuracy being lowest for squares near the middle of the distribution. Another clear pattern is that when participants misidentified a square, they tended to name the next smaller square. A Bayesian ANOVA with participant as a random effect confirmed that identification accuracy changed across the eight blocks of the learning phase ($BF_{10} = 3.92 \times 10^{18}$). The linear trend in the means across blocks was positive (95% BCI [0.17, 0.25]) while the quadratic trend was negative (95% BCI [-0.14, -0.06]), indicating that accuracy quickly increased before leveling off (means: 0.37, 0.48, 0.51, 0.59, 0.60, 0.63, 0.63, 0.62). Seven participants had accuracies that were lower than 50% in the last four blocks of the learning phase; however, none of our substantive findings for the name-based reproduction task discussed below were changed when these participants were excluded from the analyses, and so we kept them in the analyses.

Reproduction bias from long-term memory

Figure 4 presents mean reproduction biases from the name-based reproduction task segregated by distribution and anchor (both axes were scaled down by a factor of 10 to agree with our modeling analysis). The pattern of bias in Fig. 4 is markedly different from the assimilative pattern observed in Experiment 1: Squares in the positively skewed distribution were reproduced larger than squares in the negatively skewed distribution (a contrast effect), with the greatest difference being for those squares near the middle of each distribution (a quadratic effect predicted by RFT). The pattern of bias shown in Fig. 4 is consistent with the idea that stimuli are recalled based on their percentile ranks within their contextual

Table 1 Identification accuracy in the learning phase of Experiment 2: Presented versus named squares

Presented square	Size (pixels)	Named square								
		A	B	C	D	E	F	G	H	I
Positive skew										
A	40	**0.90**	0.06	0.01	0.00	0.01	0.00	0.00	0.00	0.01
B	50	0.22	**0.55**	0.18	0.01	0.02	0.00	0.00	0.00	0.00
C	60	0.04	0.17	**0.56**	0.17	0.03	0.01	0.00	0.00	0.00
D	70	0.00	0.04	0.32	**0.42**	0.12	0.07	0.01	0.01	0.00
E	80	0.00	0.04	0.11	0.39	**0.29**	0.10	0.07	0.00	0.00
F	90	0.00	0.02	0.03	0.20	0.26	**0.24**	0.20	0.04	0.01
G	100	0.00	0.02	0.01	0.08	0.17	0.23	**0.41**	0.06	0.02
H	130	0.00	0.01	0.00	0.03	0.00	0.04	0.21	**0.58**	0.13
I	160	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.12	**0.85**
Negative skew										
A	40	**0.95**	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B	70	0.25	**0.62**	0.11	0.00	0.00	0.00	0.00	0.00	0.01
C	100	0.04	0.19	**0.56**	0.14	0.05	0.01	0.00	0.00	0.00
D	110	0.01	0.11	0.21	**0.48**	0.12	0.05	0.01	0.01	0.00
E	120	0.00	0.03	0.05	0.31	**0.37**	0.17	0.05	0.01	0.01
F	130	0.00	0.02	0.05	0.15	0.27	**0.35**	0.12	0.02	0.01
G	140	0.00	0.01	0.02	0.03	0.16	0.28	**0.40**	0.07	0.02
H	150	0.00	0.01	0.01	0.03	0.01	0.04	0.25	**0.56**	0.08
I	160	0.00	0.00	0.00	0.01	0.00	0.01	0.02	0.08	**0.87**

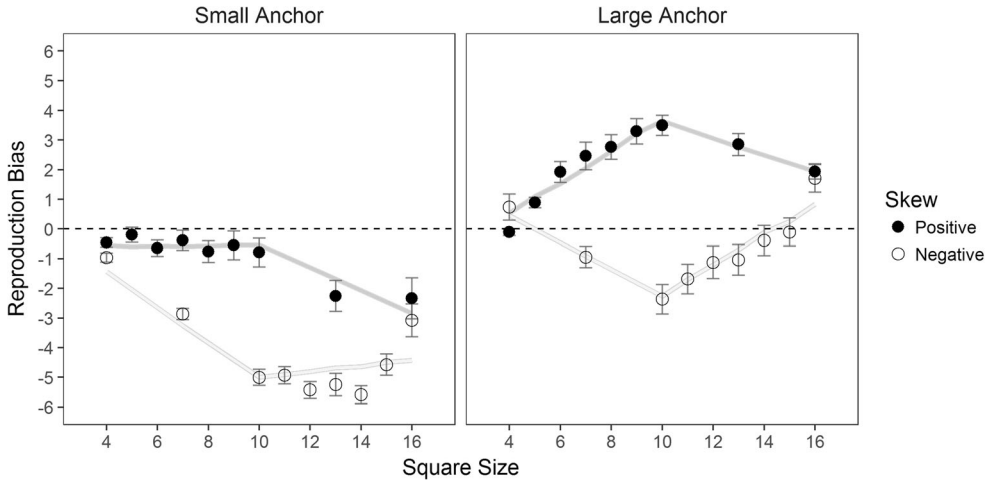

Fig. 4 Mean reproduction bias as a function of skew, anchor, and stimulus size (name-based reproduction task, Experiment 2). Lines show the fit of the modified RF model (Eq. 6). On each trial, the model prediction is the expectation of the posterior predictive distribution, given the covariates for that trial. Both axes were scaled down by a factor of 10 to match the modeling analysis. Error bars represent one standard error of the mean

distribution. There were once again large anchor effects, with reproduced squares being larger when the starting square was large compared to when it was small.

Bayesian ANOVA A Skew × Anchor × Stimulus Bayesian repeated-measures ANOVA was conducted on mean bias scores for the five squares shared by both skewing conditions (40, 70, 100, 130, and 160 pixels). There was extreme evidence for the Skew, Anchor, and Stimulus main effects (all $BF_{10} > 6.29 \times 10^{25}$), as well as the Anchor × Stimulus interaction ($BF_{10} = 1.82 \times 10^{14}$) and Skew × Stimulus interaction ($BF_{10} = 3.07 \times 10^{26}$). All of these effects were qualified by moderate evidence for the three-way interaction ($BF_{10} = 6.89$), which we probed by comparing the Skew × Stimulus interaction for each anchor separately.

As shown in Fig. 4 for both anchors, the mean biases in the positively skewed condition had a negative quadratic trend across square sizes (small anchor: 95% BCI [-10.57, 1.12]; large anchor: 95% BCI [-27.70, -15.96]), while the means in the negatively skewed condition had a positive quadratic trend across square sizes (small anchor: 95% BCI [21.47, 33.68]; large anchor: 95% BCI [22.89, 34.67]). The result is that the largest separation between the mean biases for the two skewing conditions was for the intermediate stimuli in each distribution. The Anchor × Skew × Stimulus interaction was due to the difference between the opposing quadratic trends in the large anchor condition being larger than the difference between the opposing quadratic trends in the small anchor condition (95% BCI for the difference [6.53, 30.36]), resulting in greater separation between the means for intermediate stimuli in the large anchor condition. There was also a noticeable downward slope in the small anchor means compared to the slightly positive slope for the large anchor means, reflected in the Anchor × Stimulus interaction (Fig. 4).

Computational modeling We used a modification of RFT to characterize the observed pattern of reproduction bias from long-term memory. When applied to judgment data, the fundamental assumption of the model is that an individual *i*'s internal judgment of a stimulus *j* in context *k* is a weighted combination of the range value (R_{jk}) and the frequency value (F_{jk}) for that stimulus:

$$J_{ijk} = w_i R_{jk} + (1-w_i) F_{jk}. \quad (5)$$

The range value is calculated as the proportion of the overall range of stimulus values in context *k* that fall below the value of stimulus *j*, while the frequency value is calculated as the proportion of the stimulus ranks in context *k* that fall below the rank of stimulus *j*. RFT predicts that a stimulus with an intermediate magnitude will have a higher judged magnitude when it comes from a positively skewed distribution than when the same stimulus comes from a negatively skewed distribution. When the distributions have the same range, this effect is due to the percentile rank of an intermediate stimulus being greater in the positively than in the negatively skewed distribution (Parducci, 1995).

RFT can be extended to account for reproduction bias from long-term memory. To do this, we assumed that participants were more likely to recall their implicit judgments of the stimuli generated at encoding, which are based on range-frequency values, rather than the stimulus values themselves (see Choplin & Wedell, 2014; Higgins & Lurie, 1983). Since the smallest and largest stimuli were the same in both skewing conditions, the basic RFT would predict a contrast effect for moderate stimuli such that they would be remembered as larger in the context of a positively skewed distribution than in the context of a negatively skewed distribution.

We also allowed for the possibility of assimilation effects such as those observed in Experiment 1. Three types of assimilation effects were considered: (1) global assimilation to the mean of the contextual distribution (M_k), (2) local assimilation to the anchor stimulus on each trial (AS_t), and (3) local assimilation to the *reproduced* stimulus on the previous trial (S^*_{t-1}). Because participants only saw the squares they reproduced in the name-based reproduction task, we could not include the stimulus *presented* on the previous trial, as we did in Experiment 1. Combining assimilation to the distributional mean with range-frequency weighting can result in a contrast effect for intermediate stimuli and an assimilation effect for end stimuli (Choplin & Wedell, 2014). Thus, our hybrid RF model can be formalized as follows:

$$B_{ijkt} = c_i + w_{M,i}M_k + w_{AS,i}AS_t + w_{PREV,i}S^*_{t-1} + b_i\left[w_iR_{jk} + (1-w_i)F_{jk}\right] - S_j. \quad (6)$$

In Eq. 6, c_i is an individual-level constant and b_i is an individual-level weighting of the range-frequency terms. The parameter w_i ($0 \le w_i \le 1$) determines individual i's relative weighting of range values compared to frequency values. The range value for S_j was computed as (S_j – Min)/(Max – Min), where Min and Max are the minimum and maximum of the contextual distribution, and the frequency value for S_j was computed as [rank(S_j) – 1]/(N – 1), where N is the total number of stimuli in the contextual distribution. In this experiment, Min = 4, Max = 16, and N = 9.[5] Note that removing the terms for the anchor stimulus and the previous reproduced stimulus in Eq. 6 would result in the modified RF model presented in Choplin and Wedell (2014).

We built the model as a hierarchical Bayesian model in Stan (Carpenter et al., 2016). As in Experiment 1, individual-level parameters were assumed to be drawn from group-level normal distributions (using a non-centered parameterization), and group-level means and standard deviations were given vague normal and half-normal priors, respectively. We ran four HMC chains for 2,500 samples each and discarded the first 1,250 samples from each chain as warm-up, yielding a total of 5,000 samples from the model's posterior distribution. The $\hat{R}$ statistics for each parameter were all below 1.01, indicating that the chains had sufficiently mixed and converged to the posterior distribution.

The model was able to fit the group-averaged data very well, only slightly missing the mean biases for end stimuli (Fig. 4). The negative quadratic trends for the positive skew condition and the positive quadratic trend for the negative skew condition were well captured by the high weighting of frequency values. This can be seen in the posterior distribution for the group-level mean of 1 – w (Fig. 5), which has most of its density piled up near 1.0 (95% BCI [0.94, 1.00]). Thus, the model accounts for the larger contrast effect for intermediate squares by assuming that individuals use percentile ranks when estimating sizes from memory. The posterior for the group-level mean of b, the weighting of range-frequency terms, was located far above zero (95% BCI [10.47, 11.94]), indicating that ranks strongly affected recalled magnitudes.

There was strong evidence for additional sources of bias at the group level that could not be accounted for by rank information alone. The strength of these additional sources of bias can be gauged by examining the posteriors for the group-level means of their respective weights (Fig. 5). Interestingly, there was an overall tendency to bias estimates *away* from the mean of the contextual distribution, as indicated by the posterior for the mean of the w_M parameter falling almost entirely below zero (95% BCI [-0.43, -0.03]). This global contrast effect was unexpected given that previous work had shown assimilation effects for the smallest and largest stimuli and attributed this to an assimilative bias toward the contextual mean (Choplin & Wedell, 2014). There were also group-level tendencies to assimilate toward the size of the anchor square on each trial (95% BCI [0.05, 0.11]) and toward the reproduced square size on the previous trial (95% BCI [0.03, 0.07]). Thus, our model revealed multiple sources of additional estimation bias even in the presence of strong range-frequency effects.

In addition to analyzing posteriors for the overall group means of RF model parameters, we used Bayesian ANOVAs to compare the posterior means for individual participants' parameters between skew and anchor conditions. There was extreme evidence that the parameter governing the weighting of range-frequency terms (b) differed between anchor conditions ($BF_{10} = 1.02 \times 10^7$). The b parameter controls the degree to which estimates change in response to changes in range-frequency values. On average, participants in the large anchor condition had higher b parameters than those in the small anchor condition (95% BCI for the difference [2.41, 4.33]), which can explain why the mean reproduction biases across square sizes had a more positive slope in the large anchor condition compared to the small anchor condition (Fig. 4).

Reproduction bias from short-term memory

Figure 6 shows mean reproduction biases from the immediate reproduction task as a function of distribution and anchor. The pattern of bias is quite distinct from the pattern observed when squares were recalled from long-term memory following a name cue. Importantly, squares in the positively skewed distribution were *not* reproduced larger than the same squares in the negatively skewed distribution. The pattern is also somewhat different to the pattern observed in Experiment 1, even though both tasks involved immediate reproduction from short-term memory. Here, there were few to no effects of distribution or anchor when the data were averaged over trials,

[5] Here again we scaled all stimulus values down by a factor of 10 to aid HMC sampling efficiency.

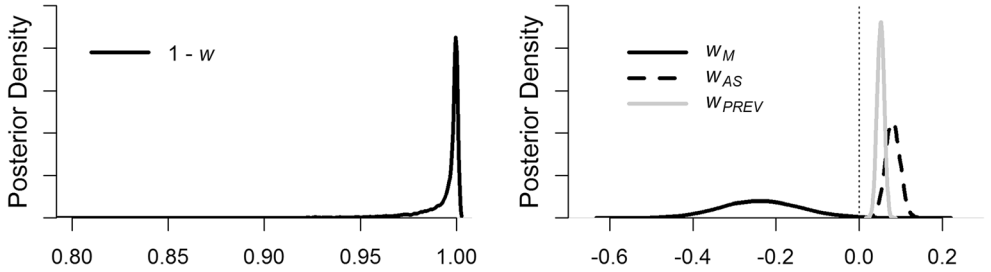

Fig. 5 **Left panel:** Posterior for the group mean of the frequency weighting parameter, 1- *w*. **Right panel:** Posteriors for the group means of w_M, w_{AS}, and w_{PREV}, i.e. the weights given to the distributional mean, anchor stimulus, and previous response in the category prototype

although there was a trend toward a skew effect for the large anchor condition. There was still a strong tendency to overestimate the smallest squares and underestimate the largest squares in each distribution.

Bayesian ANOVA A Skew × Anchor × Stimulus Bayesian repeated-measures ANOVA was conducted on mean bias scores for the five squares shared by both skewing conditions (40, 70, 100, 130, and 160 pixels). There was extreme evidence in favor of the Stimulus main effect (BF_{10} = 2.17 × 10^{33}). This effect was driven by a negative linear trend (95% BCI [-18.80, -14.47]), reflecting the overestimation of smaller squares and underestimation of larger squares, as well as a positive cubic trend (95% BCI [1.37, 5.62]), reflecting a slight diminishing of bias near end stimuli. In contrast to the results of Experiment 1, there was no evidence in favor of an Anchor effect (BF_{10} = 0.99) nor a Skew effect (BF_{10} = 0.29).

Computational modeling The same modified CA model that was tested in Experiment 1 was fit to trial-by-trial data using the hierarchical Bayesian method. The one difference is that the midpoint of the stimulus distribution in the equation for the fine-grain weighting parameter λ was 10 instead of 8 (see Eq. 3). The $\hat{R}$ statistics for each parameter were all below 1.01, suggesting that the chains had sufficiently mixed and converged to the posterior distribution.

The hierarchical model provided an adequate fit to the group-averaged data, although there was considerable misfit to the biases for smallest and largest squares in the large anchor condition (Fig. 6). The posteriors for the group-level means of the key parameters revealed that the previous stimulus once again produced a strong effect (95% BCI [0.30, 0.47]), but the running mean also contributed significantly to the category prototype (95% BCI [0.19, 0.48]). The effect of start anchors was in the same direction as in Experiment 1 but substantially reduced (95% BCI [-0.04, 0.15]) (Fig. 7). In summary, hierarchical cognitive modeling revealed that while distributional biasing effects were not apparent when collapsing the data across individuals and trials, they nevertheless were operating on a trial-by-trial basis. The posterior means for the fine-grain weighting parameter λ implied that discrimination was once again greatest near the end stimuli (Fig. 7).

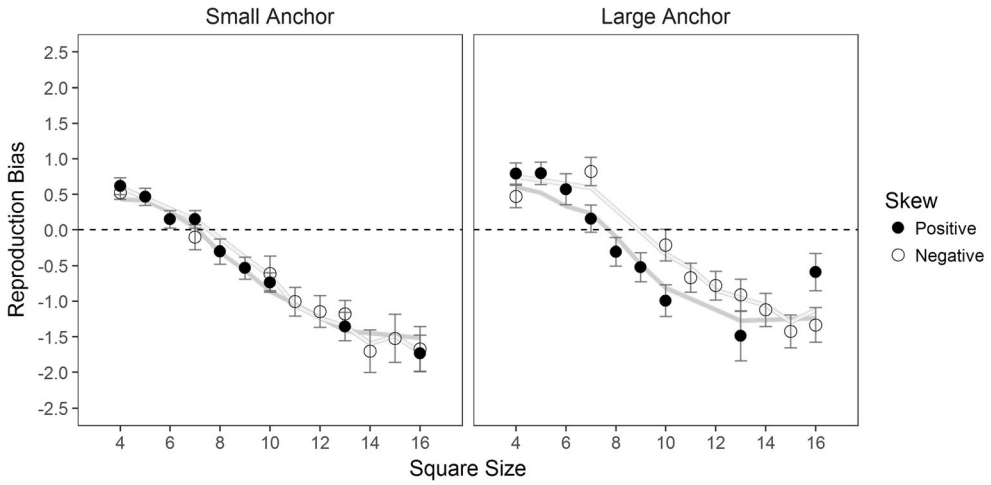

Fig. 6 Mean reproduction bias as a function of skew, anchor, and stimulus size (immediate reproduction task, Experiment 2). Lines show the fit of the modified CA model (Eq. 2). On each trial, the model prediction is the expectation of the posterior predictive distribution, given the covariates for that trial. Both axes were scaled down by a factor of 10 to match the modeling analysis. Error bars represent one standard error of the mean

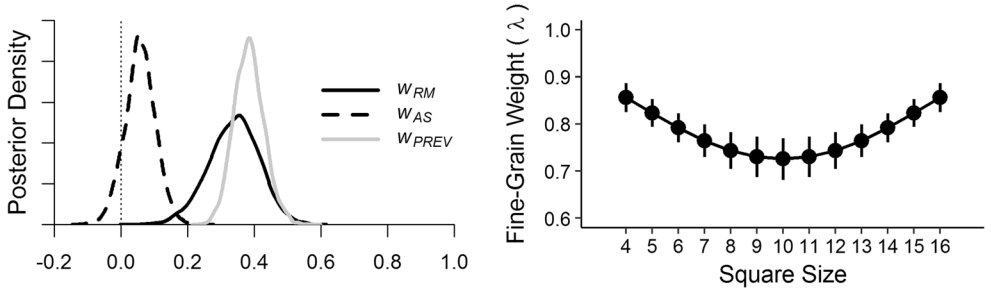

Fig. 7 **Left panel:** Posteriors for the group means of w_{RM}, w_{AS}, and w_{PREV}, i.e. the weights given to the running mean, anchor stimulus, and previous stimulus in the category prototype. **Right panel:** Posterior means for the fine-grain weighting parameter across different square sizes. Bars represent 95% Bayesian credible intervals

Discussion

The results of Experiment 2 demonstrate that the direction of bias in a stimulus reproduction task critically depends on whether the task requires retrieval from short-term or long-term memory. When squares were briefly presented on screen and then reproduced following a 1-s delay, the pattern of bias was consistent with predictions from the CA model (Crawford, 2019; Duffy et al., 2010). In particular, reproduced squares assimilated toward the running mean of the contextual distribution and toward the size of the stimulus that was presented on the previous trial. When participants were given name cues and attempted to reproduce the squares that they had previously learned to associate with those cues, a very different pattern of bias emerged. Under these conditions, target squares were reproduced larger when they were embedded in a distribution of predominantly smaller squares and vice versa. This pattern of bias was consistent with our modified RF model in which participants utilize rank-based encoding during learning and then use these values when reproducing magnitudes from name cues. These results are in line with previous research for recalling numerical values from memory (Choplin & Wedell, 2014), as well as with the general finding that rank-based encoding is prevalent when magnitudes must be held in memory for later evaluation or comparison (Hicklin & Wedell, 2007; Pettibone & Wedell, 2007; Wedell, 1996).

Computational modeling was leveraged to describe the cognitive mechanisms that may have generated the overall pattern of reproduction bias in both long-term and short-term memory tasks. A modified RF model was used to describe reproduction bias from long-term memory, while a modified version of the CA model was used to characterize reproduction bias from short-term visual memory (as in Experiment 1). Our results for the RF model were similar to previous findings from Choplin and Wedell (2014). Their experiments on cued recall of calories produced a similar pattern of bias to that shown in Fig. 4, although without anchor effects since participants simply wrote down the number of recalled calories without reference to an anchor. Their data were successfully fit with both a modified RF model and the comparison-induced distortion (CID) model (Choplin, 2007; Choplin & Hummel, 2002). The CA model was unable to account for their data, implying that assimilation toward a category prototype may not be an accurate predictor of bias effects observed when values are recalled from long-term memory.

Our results extended the findings from Choplin and Wedell (2014) from recall of discrete numerical stimuli (i.e., calorie amounts) to reproduction of analog visual stimuli. By expanding the modified RF model presented in Choplin and Wedell (2014), we were also able to demonstrate consistent assimilation toward the anchor stimuli on each trial as well as assimilation toward the size of the reproduced square on the previous trial. That contrast effects were observed at the level of the overall stimulus distribution, driven by rank-based encoding, while assimilation effects were observed at the local level, driven by recent responses and start-anchors, is consistent with previous demonstrations of opposing directions of bias that may arise from global and local contexts (e.g., Wedell, Parducci, and Geiselman, 1987).

Whereas Choplin and Wedell (2014) reported a pattern of assimilation for end stimuli and contrast for intermediate stimuli, the cued reproduction task of Experiment 2 did not result in assimilation at the end points. In the earlier research, the assimilation effects were modeled by a positive weighting of the category mean within the modified RF model or as a natural consequence of the CID model. The failure to find this effect in the current experiment means that the data are difficult to explain within the CID framework, which entails this reversal of context effects. The modified RF model in the present study found a significant negative weight for the distribution mean rather than a positive weight. It is unclear what procedural differences between the two experiments can explain this discrepancy. It may be due to differences in stimuli (analog or numerical), learning procedures, response procedures, or other design elements.

Our results for the immediate reproduction task were largely consistent with our results from Experiment 1, although the assimilation effects predicted by the CA model were not reliable when the data were averaged across trials (Fig. 6). The advantage of using hierarchical computational modeling to

describe the data at the trial-by-trial level is that it can detect effects that may be obscured when the data are aggregated across trials and individuals. The effects present in averaged data do not always represent the effects present in a specific individual's data (Estes, 1956). Although there was very little separation between the means for each skewing condition, our implementation of the CA model revealed systematic bias toward the running mean of the stimulus distribution and the stimulus presented on the previous trial, which agrees with recent reexaminations of the data from Duffy et al.'s (2010) experiment (Crawford, 2019; Duffy & Smith, 2018). We were unable to replicate the strong anchor effects found for the short-term reproduction task in Experiment 1, although here the anchors were varied between participants. The lack of robust start anchor effects in the short-term memory retrieval portion of Experiment 2 is perplexing given very large effects were found for the same participants when engaged in the long-term memory reproduction task.

General discussion

Stimulus reproduction tasks reveal characteristic patterns of bias that can shed light on underlying memory encoding and retrieval processes. The aim of the current study was to examine how two patterns of reproduction bias – (1) assimilation toward the central tendency, or category prototype, of the contextual distribution and (2) contrast effects driven by rank-based encoding and retrieval – might depend on whether the reproduction task requires retrieval from short-term or long-term memory.

Consistent with prior work, we demonstrated that reproduced square sizes assimilate toward the mean of the contextual distribution when the task requires retrieval from short-term visual memory (Crawford, 2019; Duffy et al., 2010). We modeled this effect using a modified version of the CA model in which the category prototype was comprised of the running mean and the prior stimulus value, both of which predict skewing, along with the start anchor value, which does not predict skewing. The biasing effects of the running mean, and possibly the prior stimulus, can be justified from a Bayesian perspective in that these tend to bias estimates toward the prototypical value and thereby reduce error variance when there is uncertainty about the true stimulus value (Duffy et al., 2010; Huttenlocher et al., 2000). This finding is largely in agreement with recent re-examinations of the CA theory (Crawford, 2019; Duffy & Smith, 2018). However, the strong bias toward the start-anchor (Experiment 1) does not minimize error variance and appears to reflect a tendency to use salient cues to anchor judgments even when these are not representative of the contextual distribution. One way to conceive of these results is that the category prototype may reflect multifaceted sources of information, some more representative than others of the category exemplars.

Experiment 2 demonstrated that cued retrieval of stimulus magnitudes from long-term memory yields a very different pattern of bias that reflects contrast rather than assimilation effects. These effects were modeled as reflecting rank-based encoding of stimulus magnitudes that were used at retrieval to reconstruct stimulus values. The model also revealed a negative weight of the distributional mean (adding to contrast effects) as well as positive weights of the start anchor and prior response. Thus, our findings suggest that stimulus reproduction from long-term memory involves a more complicated recruitment of multiple sources of contextual information than reproduction from short-term memory, resulting in simultaneous assimilation and contrast effects at the local and global levels, respectively.

An important direction for future research concerns the factors that determine the sources and directions of bias when recalling values from long-term memory. In Experiment 2, the learning task mapped category names directly onto stimulus magnitudes, as these were unidimensional stimuli. Thus, the name-learning task based on differences in magnitudes may have made rank values more salient and facilitated their use in estimation. However, with multidimensional stimuli, the salient dimension can be uncorrelated with the dimension being recalled, as when different colored squares are mapped onto name cues that are then later used to recall another dimension such as size. Prior research using rating-scale responses has shown that the encoding environment is a key determinant of the observed context effects (Hicklin & Wedell, 2007; Pettibone & Wedell, 2007). Given that events and objects in the real world are multidimensional and encoded into long-term memory in a variety of ways, how bias depends on the relationship between encoding and retrieval contexts is critical to understanding applications of these results in the real world.

Data Availability The datasets analyzed during the current study and the code for reproducing all analyses are available in the Open Science Framework (OSF) repository, https://osf.io/8abj4/.

Appendix

The size ratings in Experiment 1 were fit with a modified RF model, similar to the model used to fit reproduction bias from long-term memory in Experiment 2 (Eq. 6). Because squares were encountered sequentially in the rating task (as opposed to in a prior learning phase, as in Experiment 2), we computed range and frequency values on trial t for each participant based on the squares that they encountered up to that trial. That is, the range value for square j encountered on trial t (R_{jt}) was computed as $(S_j - \mathrm{Min}_t)/(\mathrm{Max}_t - \mathrm{Min}_t)$, where Min_t and Max_t are the minimum and maximum square sizes encountered up

to trial t, and the frequency value was computed as $[\text{rank}_t(S_j) - 1]/(N_t - 1)$, where the rank is based on the set of N squares encountered up to trial t.

The model prediction for individual i's judgment of the j^{th} square on trial t was computed with the following equation:

$$J_{ijt} = c_i + w_{RM,i} RM_t + w_{PREV,i} S_{t-1} + b_i \left[w_i R_{jt} + (1 - w_i) F_{jt} \right].$$

In this equation, c_i is an individual-level constant and b_i is an individual-level weighting of the range-frequency terms. The parameter w_i ($0 \le w_i \le 1$) determines individual i's relative weighting of range values compared to frequency values. In addition to range and frequency values, the model assumed that the rating for the square on trial t could be biased by the running mean of square sizes encountered up to trial t (RM_t) and the size of the square rated on the previous trial (S_{t-1}). The effects of these additional sources of bias were also allowed to vary across individuals. Unlike in Eq. 6, there was no term for anchor bias in this model because anchors were not present in the rating task.

The model was estimated in a hierarchical Bayesian fashion using the same procedures and diagnostic criteria discussed in the main text. The posterior distribution for the group mean of the frequency weighting parameter $1 - w$ revealed that participants placed substantial weight on percentile ranks when rating square sizes (95% BCI [0.38, 0.52]). This parameter accounts for the curvature in the mean ratings, with moderate squares being judged larger when embedded in a distribution of predominantly smaller squares than in a distribution of predominantly larger squares (Fig. 3). The weighting of ranks close to the 0.50 value is typically found in such tasks (Parducci, 1995); however, it was not nearly as high as that observed in Experiment 2, when participants attempted to reproduce squares from long-term memory (compare Figs. 4 and 5). We also found that ratings were biased in the direction of the running mean (w_{RM} 95% BCI [0.21, 0.40]) and the size of the square on the previous trial (w_{PREV} 95% BCI [0.01, 0.02]). Thus, while stimulus ranks produced contrast effects on category ratings, in line with RFT (Parducci, 1995), they also assimilated to some degree toward the central tendency of the contextual distribution and toward recently encountered stimuli.

References

Betancourt, M. J., & Girolami, M. (2015). Hamiltonian Monte Carlo for Hierarchical Models. In S.K. Upadhyay, U. Singh, D.K. Dey, A. Loganathan (Eds.), *Current Trends in Bayesian Methodology with Applications*. (pp. 79-100). CRC Press.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology, 66,* 5-20.

Biernat, M., & Manis, M. (2007). Stereotypes and Shifting Standards: Assimilation and Contrast in Social Judgment. In D. A. Stapel & J. Suls (Eds.), *Assimilation and contrast in social psychology* (pp. 75-97). New York, NY, US: Psychology Press.

Biernat, M., Kobrynowicz, D., & Weber, D. L. (2003). Stereotypes and shifting standards: Some paradoxical effects of cognitive load. *Journal of Applied Social Psychology, 33,* 2060-2079.

Carpenter, B., Gelman, A., Hoffman, M., & Lee, D. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software, 76.* https://doi.org/10.18637/jss.v076.i01.

Choplin, J. M. (2007). Toward a comparison-induced distortion theory of judgment and decision making. In J. A. Elsworth (Ed.), *Psychology of decision making in education, behavior and high risk situations* (pp. 11-40). Hauppauge, NY: Nova Science.

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

Choplin, J. M., & Wedell, D. H. (2014). How many calories were in those hamburgers again? Distribution density biases recall of attribute values. *Judgment and Decision Making, 9,* 243-258.

Cornsweet, T. N. (1962). The staircase- method in psychophysics psychophysics. *Amencan American Journal of Psychology, 75,* 485-491.

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

Crawford, L. E., Huttenlocher, J., & Engebretson, P. H. (2000). Category effects on estimates of stimuli: Perception or reconstruction? *Psychological Science*, *11*, 280-281.

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

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.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic. *Psychological Science*, *17*, 311-318.

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

Fitting, S., Allen, G. L., & Wedell, D. H. (2007). Remembering places in space: A human analog study of the Morris water maze. In T. Barkowsky, M. Knauff, G. Ligozat, & D. R. Montello (Eds.), *Spatial cognition V: Reasoning, action, interaction*, *LNAI 4387* (pp. 59–75). Berlin: Springer.

Fitting, S., Wedell, D. H., & Allen, G. L. (2007). Memory for spatial location: Cue effects as a function of field rotation. *Memory and Cognition, 35,* 1641-1658.

Fitting, S., Wedell, D. H., & Allen, G.L. (2008). Cue usage in memory for location when orientation is fixed. *Memory and Cognition, 36,* 1196-1216.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton: CRC Press.

Haun, D. B. M., Allen, G. L., Wedell, D. H. (2005). Bias in spatial memory: A categorical endorsement. *Acta Psychologica, 118*, 149-170.

Helson, H. (1964). *Adaptation-level theory.* New York: Harper & Row.

Helström, A, (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin*, *97*, 35-61.

Hicklin, S. K., & Wedell, D. H. (2007). Learning group differences: Implications for contrast and assimilation in stereotyping. *Social Cognition, 25,* 410-454.

Higgins, E. T., & Lurie, L. (1983). Context, categorization, and recall: The "change-of-standard" effect. *Cognitive Psychology, 15,* 525-547.

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology,*

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

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, *98*, 352-376.

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

Jones, M. R., & McAuley, J. D. (2005). Time judgments in global temporal contexts. *Perception and Psychophysics*, *67,* 398-417.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology, 55,* 1-7.

Marks, L. E. (1992). The slippery context effect in psychophysics: Intensive, extensive and qualitative continua. *Perception and Psychophysics*, *51*, 187-198.

Morey, R. D., & Rouder, J. N. (2014). *BayesFactor 0.9.6. Comprehensive R Archive Network.* Retrieved from http://cran.r-project.org/web/packages/BayesFactor/index.html

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, *35*, 136–164.

Parducci, A. (1995). *Happiness, pleasure and judgment: The contextual theory and its applications*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Pettibone, J. C., & Wedell, D. H. (2007). Of gnomes and leprechauns: The recruitment of recent and categorical contexts in social judgment. *Acta Psychologica, 125,* 361-389.

Poulton, E. C., Simmonds, C. V., & Warren, R. M. (1968), Response bias in very first judgments of reflectance of grays: Numerical versus linear estimates. *Perception and Psychophysics, 3,* 112-114.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods, 22,* 304-321.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*, 1-75.

Ryan, L. J. (2011). Temporal context affects duration reproduction. *Journal of Cognitive Psychology*, *23,* 157-170

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

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32,* 1248-1284.

Stan Development Team (2016). RStan: The R interface to Stan. Retrieved from http://mc-stan.org.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53*, 1-26.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1130.

Ungemach, C., Stewart, N., & Reimers, S. (2011). How incidental values from our environment affect decisions about money, risk, and delay. *Psychological Science, 22,* 523-560.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25,* 58-76.

Ward, L. M. (1979). Stimulus information and sequential dependencies in magnitude estimation and cross-modal matching. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 444-459.

Wedell, D. H. (1996). A constructive-associative model of the contextual dependence of unidimensional similarity. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 634-661.

Wedell, D. H. (2008). A similarity-based range-frequency model for two category rating data. *Psychonomic Bulletin and & Review, 15*, 638-643.

Wedell, D. H., Hicklin, S. M., & Smarandescu, L. O. (2007). Contrasting models of assimilation and contrast. In D. A. Stapel and J. Suls (Eds.) *Assimilation and Contrast in Social Psychology* (pp. 45-74), New York: Psychology Press.

Wedell, D. H., & Senter, S. M. (1997). Looking and weighting in judgment and choice. *Organizational Behavior and Human Decision Processes, 70,* 41-64.

Wedell, D. H., Parducci, A., & Geiselman, R. E. (1987). A formal analysis of ratings of physical attractiveness: Successive contrast and simultaneous assimilation. *Journal of Experimental Social Psychology, 23,* 230-249.