Journal of Experimental Social Psychology 85 (2019) 103847

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Comparing value coding models of context-dependence in social choice☆

Linda W. Chang, Samuel J. Gershman, Mina Cikara*

Department of Psychology, Harvard University, Cambridge, MA 02138, United States of America

ARTICLE INFO

Keywords:
Context-dependence
Social cognition
Decision-making
Value coding

ABSTRACT

Decision-makers consistently exhibit violations of rational choice theory when they choose among several alternatives in a set (e.g., failing to buy the best product in a set when it is presented alongside high-quality alternatives). Many of society's most significant social decisions similarly involve the joint evaluation of multiple candidates. Are social decisions subject to the same violations, and if so, what account best characterizes the nature of the violations? Across five studies, we tested whether decision-makers exhibit context-dependent preferences in hiring scenarios and past U.S. congressional race outcomes and compared different models of value coding as sources of the hypothesized context-dependence. Studies 1a, 1b, and 1d revealed that a divisive normalization value coding scheme best characterized participants' choices across a series of hiring decisions, and that participants exhibited context-dependent preferences. However, the distractor had the opposite effect of that predicted by divisive normalization once we accounted for the random effect of participant: as the value of the distractor increased, participants were *more* likely to hire the highest-valued candidate. In Study 2, we used a combination of archival electoral data and survey data to examine whether normalization models could explain the outcomes of congressional elections. Electoral outcomes were predicted by political candidates' inferred competence, but this time in line with the divisive normalization account. Our findings offer mixed support for a formal, neurobiologically-derived account of when and how specific alternatives exert their effects on social evaluation and choice, and highlight conditions under which high-value distractors increase versus decrease relative choice accuracy.

1. Introduction

Psychologists have long recognized that the construction of social choice sets affects which stereotypes or target features become most salient, and as a consequence, how each constituent person or social group within the choice set is evaluated (e.g., Biernat & Manis, 1994; Judd & Park, 1993; Oakes, Haslam, & Turner, 1998; Pan, O'Curry, & Pitts, 1995; Trope & Mackie, 1987; Wyer, Sadler, & Judd, 2002). This extends to evaluations in professional contexts (e.g., Bohnet, van Geen, & Bazerman, 2015; Highhouse, 1996; Leung & Koppman, 2018; Norton, Vandello, & Darley, 2004; Simonsohn & Gino, 2013). For example, a recent study examining choice-set dependence in hiring indicated that when a group of applicants were majority white, participants chose to hire a white candidate more often than based on chance alone; however, when the group of applicants was majority female or majority black, participants also chose a female or black candidate more often than chance (Johnson, Hekman, & Chan, 2016). Said another way: when the choice set indicated that the status quo was white and male, a white male candidate was preferred; when the status quote was non-white-male, a non-white-male candidate was preferred.

Despite choice architecture's well-documented impact on social decision-making across a variety of consequential contexts, we know relatively little about when and how specific alternatives have the effects they have on evaluation and choice (as compared to the consumer behavior domain). Here, we adopt an inter-disciplinary approach, integrating models from cognitive and social psychology, neuroeconomics, and computational neuroscience to examine a neglected, but potentially powerful explanation of context-dependence in social decision-making: normalized value coding.

1.1. Normalization accounts of context-dependence

Rational theories of choice predict that decision makers' preferences between any two options should remain the same irrespective of the number or quality of other options: a property known as independence of irrelevant alternatives (IIA; Luce, 1959; Sen, 1971). More concretely,

☆ This research was supported by a National Science Foundation Graduate Research Fellowship (DGE-1745303 awarded to LWC) and by a National Science Foundation CAREER award (BCS-1653188 awarded to MC).

* Corresponding author at: Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, United States of America.
E-mail address: mcikara@fas.harvard.edu (M. Cikara).

https://doi.org/10.1016/j.jesp.2019.103847
Received 26 January 2019; Received in revised form 15 July 2019; Accepted 15 July 2019

rational choice theory proposes that we make all pairwise comparisons of the available options (A, B, C), construct a preference hierarchy (A > B > C), and then make decisions accordingly. In this scheme, adding a third inferior option (C) should not affect preferences between A and B. Yet, humans, monkeys, birds, bees, even ameboid organisms, reliably violate this assumption (Bateson, Healy, & Hurly, 2003; Huber, Payne, & Puto, 1982; Hurly & Oseen, 1999; Latty & Beekman, 2011; Louie, Khaw, & Glimcher, 2013; Shafir, Waite, & Smith, 2002; Simonson, 1989). Specifically, people who rank two alternatives A > B, will sometimes show a preference for B over A once C is added to the choice set (Tversky, 1969). Rather than marking a defect in human decision-making machinery, recent frameworks suggest that these 'violations' arise from a selective integration mechanism which ultimately leads to better decisions given the noise intrinsic to information processing (Howes, Warren, Farmer, El-Deredy, & Lewis, 2016; Tsetsos et al., 2016). Though this phenomenon has been documented widely in consumer behavior contexts (e.g., Huber et al., 1982; Louie et al., 2013; Simonson, 1989), those studies that have examined context-dependence in the social domain have focused almost exclusively on a highly constrained choice set: one in which two options represent perfect tradeoffs on two attributes in the presence of a "decoy," which also has a very specific attribute profile (Chang & Cikara, 2018; Herne, 1997; Highhouse, 1996; Pan et al., 1995, Pettibone & Wedell, 2000; Sedikides, Ariely, & Olsen, 1999; with one exception: Furl, 2016). Of course, social choice sets very rarely conform to these parameters. Here we turn to alternative models of context-dependence, which are not bound by these constraints.

Researchers have proposed many different accounts of context-dependence. Early theories focused on higher-order cognitive mechanisms—for example, how people attend to and dynamically integrate option attributes over the course of the decision-making process (Roe, Busemeyer, & Townsend, 2001; Simonson, 1989; Turner, Schley, Muller, & Tsetsos, 2018; Tversky & Simonson, 1993; Tversky, 1969; Usher & McClelland, 2001; see Busemeyer, Gluth, Rieskamp, & Turner, 2019 for recent review of this class of models)—whereas alternative accounts suggest context-dependence emerges as a function of value coding itself (Louie et al., 2013; Noguchi & Stewart, 2018; Soltani, De Martino, & Camerer, 2012; see also theoretical predecessors: Anderson, 1971; Fechner, 1860; Mellers & Birnbaum, 1983; Parducci, 1965; Smith, Diener, & Wedell, 1989; Stevens, 1961; Wedell & Parducci, 1988). We focus on an example of the latter here: normalization.

Very broadly, normalization refers to scaling inputs by other nearby inputs to reduce redundancy in signal processing. It was proposed as a canonical computation that operates in various neural systems and was originally developed to explain non-linear responses in primary visual cortex (for review see, Carandini & Heeger, 2012). In this context, normalization refers to when the activity of a neuron is scaled by the summed activity of a large pool of neighboring neurons. Recent evidence suggests that normalization may also apply to the representation of values associated with different choice options (Khaw, Glimcher, & Louie, 2017; Louie et al., 2013; Louie, Grattan, & Glimcher, 2011; Rangel & Clithero, 2012). This value representation is encoded in normalized form where neural firing rates increase with the value of the represented action and decrease with the value of alternative actions. Under this account, neural encoding is inherently context-dependent, such that the value of an action is explicitly dependent on the value of the available alternatives. Normalization models result in context-dependence because adding an irrelevant alternative alters the value of the remaining options—either increasing or decreasing the relative value difference between the original options.

Different models of normalization make different predictions about context-dependent choice. We describe these predictions for a set of three options whose values (when measured in isolation) are ordered as follows: target > competitor > distractor. In divisive normalization models, the value of each option is scaled by the summed value of all options in the choice set (Louie et al., 2013). Compared to untransformed values, divisive normalization predicts IIA violations are more likely when the value of the distractor is higher. Why? As the distractor within a choice set increases in value, it becomes more difficult for decision-makers to discriminate between the target and competitor because the relative value difference between them has decreased (i.e., both the target's and the distractor's values have been scaled by a larger sum). In contrast, in range-normalization models, the value of each option is scaled by the absolute difference of the highest and lowest value options in the choice set (Soltani et al., 2012). Therefore, range normalized values predict IIA violations are more likely when the value of the distractor is *lower*. No work of which we are aware has tested whether such neurobiologically-derived models best account for social decision-making in consequential contexts. Furthermore, this approach has the benefit of generalizing beyond decisions in which the options must represent tradeoffs on no more than two attributes (as in decoy effects).

1.2. Overview of the current studies

In Studies 1a – 1d we test (i) which model of value coding best captures participants' empirical choice patterns, and (ii) whether participants' choices are subject to IIA violations in a hypothetical hiring scenario. In Study 2, we use a combination of archival electoral data and survey data to test whether normalization models can explain the final outcome of three-way U.S. congressional elections. We report all measures, manipulations, and exclusions in these studies.

2. Studies 1a – 1d

Participants in Study 1a completed two phases (based on Louie et al., 2013): first they reported how likely they would be to hire each of 30 candidates, twice; then they made hiring decisions across a series of

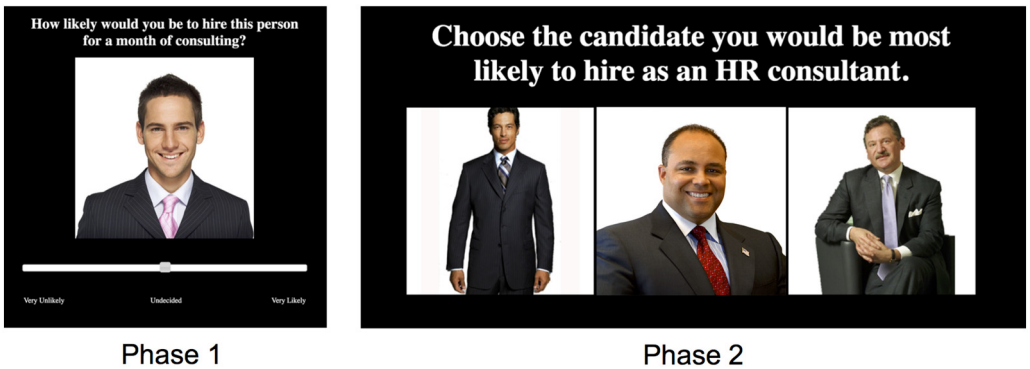

Fig. 1. Studies 1a – 1d: Example images from Phase 1 and Phase 2. Phase 1: Participants made initial evaluations for 30 individual candidates, twice. Phase 2: Participants completed 250 trinary-choice hiring trials. Each trial consists of three candidates: target, competitor, and distractor.

trinary choice sets (i.e., an array of three candidates; see Fig. 1). Each trinary choice set in Phase 2 was based on participants' own ratings from Phase 1 and consisted of two high-value candidates—a target (highest-valued) and a competitor (second highest-valued)—plus one distractor. Participants in Study 1b did the same task, except that candidates were presented sequentially in Phase 2 to ensure participants had encoded all of the options. We then did a formal model comparison to determine whether no-, range-, or divisive -normalization best characterized participants' choices.

3. Methods

3.1. Study 1a

3.1.1. Participants and exclusions

We aimed for a minimum of 40 participants after exclusions (based on Louie et al., 2013), because this is a massively repeated-measures design (sample determined prior to data collection). We recruited 44 participants from the undergraduate study pool, who completed the study for course credit. Of these, 4 participants were excluded due to computer malfunctions in the experimental session. A further 4 participants were excluded for incoherent responses within the study (2 for low correlations between their two Phase 1 hireability ratings, and 2 for heavily skewed distributions of mean Phase 1 hireability ratings). This resulted in a final sample size of $N = 36$ participants (14 female, 22 male; $M_{age} = 19.61$ years, $SD = 1.29$).

3.1.2. Procedure. Phase 1

We asked participants to imagine that they had been put in charge of hiring a human resources consultant to advise their boss on strategy and that for each candidate they should indicate how likely they would be to hire each person. We randomly sampled candidate photos from a larger photoset (taken in part from Cikara & Fiske, 2011) and presented them in randomized order. Participants made initial evaluations for all 30 candidate photos once, and then made the same evaluations again for a second time. This allowed us to screen participants' responses for inconsistency and to compute a more stable hireability value (i.e., the mean of the two ratings) for each candidate. In each evaluation trial, participants viewed an image of a candidate on a computer screen with the question, "How likely would you be to hire this person for a month of consulting?" and responded using a mouse-controlled slider bar (0–100; *very unlikely* to *very likely*, though the number associated with their response did not appear on the screen).

3.1.3. Participant-specific choice set construction for phase 2

We programmed the task so that candidates were automatically sorted by their mean hireability values from Phase 1 into two groups: a target group (10 highest-ranked) and a distractor group (20 lowest-ranked). We constructed the target pairs in the trinary-choice trials by taking 25 of the 45 possible combinations of the 10 identified targets. We presented each participant with 5 pairs with a difference of 1 in ranking (e.g., first vs. second ranked candidate), 4 pairs with a distance of 2 and 3 in ranking, 3 pairs with a distance of 4 and 5 in ranking, 2 pairs with a distance of 6 and 7 in ranking and 1 pair with a distance of 8 and 9 in ranking. We chose 10 distractors from the identified set of 20, using odd-ranked distractors (11, 13, …, 19) and each of these 10 distractors was presented with the 25 different target pairs, where each trinary-choice set was presented only once.

3.1.4. Phase 2

Participants completed 250 trinary-choice hiring trials. In each trial, participants viewed three candidates (target, competitor, and distractor) and used the mouse to indicate which one they would hire as an HR consultant. The location of each candidate on the screen (left, middle, or right) was randomly assigned in each trial (see Fig. 1).

3.1.5. Analyses

We fit mixed-effects logistic models in R 3.5.0 (R Core Team, 2018) using the *lme4* package (version 1.1.18.1; Bates, Maechler, Bolker, & Walker, 2015). For the untransformed model, the mean hireability ratings or 'value' of the candidates remained unchanged. We computed divisive normalization values by dividing each option value by the sum of all option values (target, competitor, distractor) in each trial, and computed range normalization values by dividing each option value by the difference between the target value and the distractor value in each trial. We compare all three models against one another: the untransformed model, where the value of each option is not affected by other options in the choice set, and the two different models of context-dependent value coding (divisive normalization, range normalization).

Sensitivity analyses for each study were conducted using Monte Carlo simulation via the *simr* package (version 1.0.4; Green & MacLeod, 2016) on the best fitting model for each dataset. Power was calculated by repeatedly drawing new values for the response variable from a distribution based on the fitted model, refitting the model, and then testing the statistical significance of a parameter. Each model in Studies 1a-1d had two parameters (difference between target and competitor, distractor value). We report post-hoc power for both parameters. Since our effect of interest is how distractor value affects choice, we focused our sensitivity analysis on this parameter. For each parameter, we ran 1000 simulations.

3.2. Study 1b

3.2.1. Participants and exclusions

As in Study 1a, we aimed for a minimum of 40 participants after exclusions (to replicate Louie et al., 2013). We recruited 46 participants from the undergraduate study pool, who completed the study for course credit or for pay. Of these, 1 participant was excluded for computer malfunctions in the experimental session. A further 8 participants were excluded for incoherent responses within the study (4 for low correlations between their two Phase 1 hireability ratings, 4 for heavily skewed or bimodal distributions of mean Phase 1 hireability ratings). This resulted in a final sample size of $N = 37$ participants (28 female, 9 male; $M_{age} = 20.16$ years, $SD = 1.32$).

3.2.2. Procedure

The procedure was identical to Study 1a, with one exception. In Phase 2, for each trinary choice set, we presented participants with each candidate in isolation for 1 s from left to right in their respective locations (left, middle, right), before presenting all three candidates jointly. Once candidates were jointly presented, participants could make a decision between the three candidates. We did this to ensure participants paid equal attention to all three options before making a choice (see recent discussion of the role of attention to the distractor in instances of IIA; Gluth, Spektor, & Rieskamp, 2018).

3.2.3. Analyses

Analyses were identical to Study 1a.

3.3. Study 1c

3.3.1. Participants and exclusions

We wanted to increase our sample size relative to the first two studies so we recruited 61 participants from the undergraduate study pool, who completed the study for course credit. Of these, 8 participants were excluded for incoherent responses within the study (6 for low correlations between their two Phase 1 hireability ratings, 2 for heavily skewed or bimodal distributions of mean Phase 1 hireability ratings). This resulted in a final sample size of $N = 53$ participants (32 female, 21 male; $M_{age} = 20.06$ years, $SD = 1.93$).

3.3.2. Procedure

Because the results of Studies 1a and 1b were the same, we reverted back to the design of Study 1a for this and the next study.

3.4. Study 1d

3.4.1. Participants and exclusions

Study 1c did not replicate 1a and 1b. As such, we conducted an even higher power replication which aimed for 150 participants after exclusions. We recruited 178 participants through the Decision Science Laboratory, who completed the study for pay. Of these, 2 participants were excluded for computer malfunctions in the experimental session. A further 26 participants were excluded for incoherent responses within the study (12 for low correlations between their two Phase 1 hireability ratings, 14 for heavily skewed or bimodal distributions of mean Phase 1 hireability ratings). This resulted in a final sample size of $N = 150$ participants (83 female, 65 male, 1 other, 1 did not wish to provide; $M_{age} = 31.17$ years, $SD = 15.04$). 92 participants were community members and 58 participants were undergraduates.

3.4.2. Procedure

The procedure was identical to Study 1a.

4. Results

We report regression results for the "winning" model for each study below, but include results for each model within each study in a table in the supplemental materials (See Fig. 2). Results for Studies 1a-1d are robust even when we remove trials with response times > 1.5 times the interquartile range from the upper (or lower) limit of the interquartile range. Here, we report results including all data, but see tables in supplemental materials for results *excluding* response time outliers.

4.1. Study 1a

Likelihood of hiring ratings were highly reliable across repetitions in Phase 1 (e.g., median within-participant correlation: $r = 0.94$) and across our sample (average $r = 0.93$ 95% CI [0.93, 0.94], $t(1078) = 86.16$, $p < .001$). The hireability ratings were also good predictors of participants' choices in Phase 2. Across the population, participants selected the target on 74.2% of trials, the competitor on 24.5% of trials, and the distractor on only 1.4% of trials. For the main analysis, we excluded the trials ($n = 122$) in which the distractor candidate was chosen (leaving $n = 8878$). As such, our analyses reflect relative (rather than absolute) choice accuracy (Gluth et al., 2018): not choosing the target constitutes a violation of the IIA.

We fit mixed-effects logistic regression models which treated probability of hiring the target candidate as the outcome variable, and the difference in value between the target and competitor and the distractor value as predictor variables, including participant as a random effect. Note that this difference score and distractor value varied by model: the untransformed model used the difference between the target and competitor values and the value of the distractor as predictors; the divisive normalization model used the difference between the target and competitor values and the value of the distractor, each scaled by the sum of all three option values in the choice set, as the predictors; the range normalization model used the difference between the target and competitor values and the value of the distractor, each scaled by the difference between the target and the distractor, as the predictors.

In order to investigate which model best captured the empirical choice patterns, we compared untransformed vs. divisive normalization vs. range normalization values as inputs, and selected the best model using the Bayesian information criterion (BIC). We found that the divisive normalization model had the lowest BIC (8886.28), followed by the untransformed model (8941.78), and the range normalization model (9123.40).

In order to examine whether naïve participants' choices were

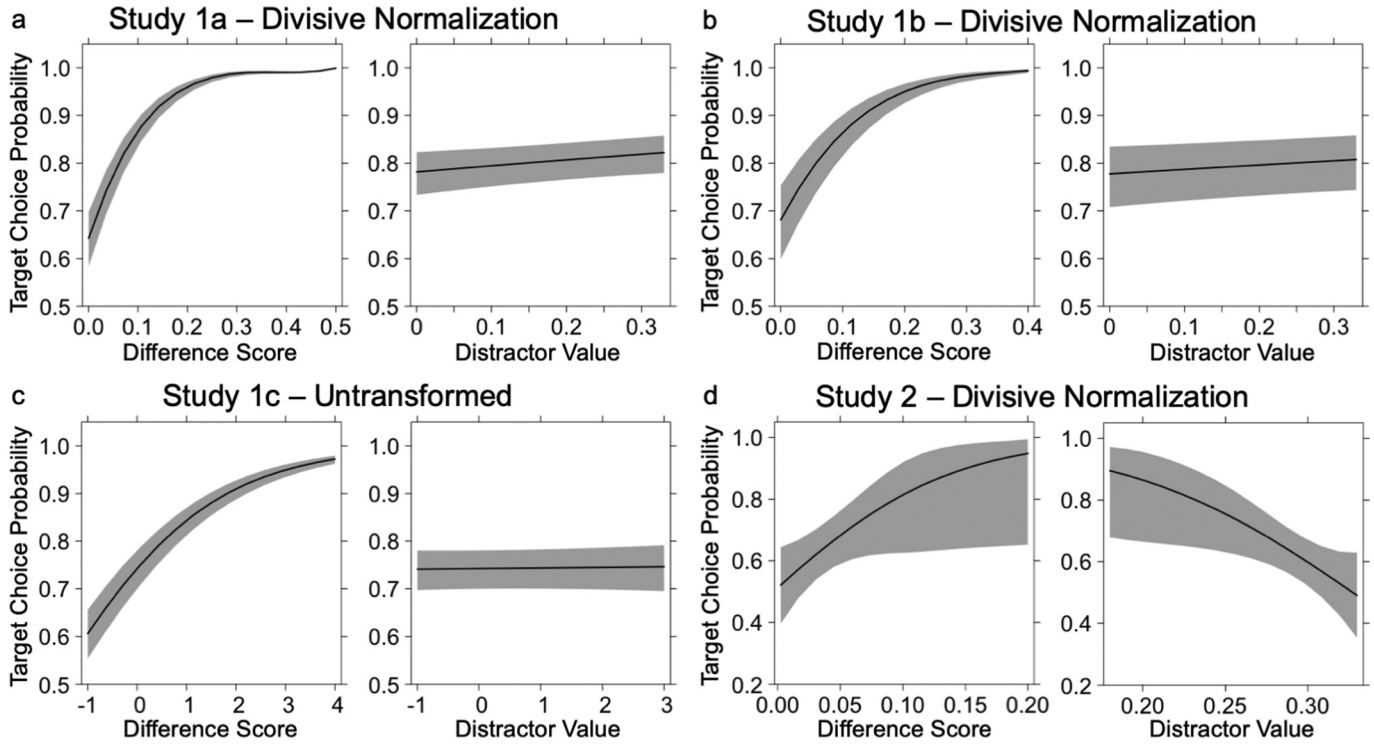

Fig. 2. Best fit mixed-effects logistic regressions for each study predicting target choice as a function of the difference between the target and competitor, and distractor value, including random effect of participant. The black lines represent model-predicted choice probabilities with error bands denoting 95% CIs. **a.** Study 1a: Divisive normalization model. **b.** Study 1b: Divisive normalization model **c**. Study 1c: Untransformed model. **d.** Study 2: Divisive normalization model.

subject to IIA violations, we examined the predictors of the divisive normalization model. As predicted by any model of value-based choice, we found that the difference in value between target and competitor significantly predicted the likelihood that the target was hired, $b = 12.91$ 95% CI [11.60, 14.25], $p < .001$. In other words, as the value difference between the target and the competitor increased, participants were more likely to select the target. Importantly, we also found that the distractor value significantly predicted the likelihood that the target was hired, $b = 0.77$ 95% CI [0.16, 1.38], $p = .013$. This suggests that while participants' choices were subject to IIA violations, these violations occur in the opposite direction of the divisive normalization model; as the value of the distractor increased, participants were more likely to hire the target.

Simulations revealed we had 66.2% 95% CI [63.17, 69.13] power to detect the effect of distractor value on choice and 100% 95% CI [99.63, 100] power to detect the effect of difference in value between target and competitor ($b = 12.90$) on choice. Sensitivity analyses revealed we would have been able to detect a minimum effect size of 0.91 for distractor value (with power = 0.8, and $\alpha = 0.05$).

4.2. Study 1b

Likelihood of hiring ratings were highly reliable across repetitions (e.g., median within-participant correlation: $r = 0.95$) and across our sample ($r = 0.94$ 95% CI [0.94, 0.95], $t(1108) = 95.74$, $p < .001$). The hireability ratings were also good predictors of participant's choices in Phase 2. Across the population, participants selected the target on 73% of trials, the competitor on 25.4% of trials, and the distractor on only 1.6% of trials. For the main analysis, we excluded the trials ($n = 150$) in which the distractor candidate was chosen (leaving $n = 9078$).

We employed the same model structure in Study 1a and again compared untransformed, divisive normalization, and range normalization values as inputs. Replicating Study 1a, we found that the divisive normalization model had the lowest BIC (9110.52), followed by the untransformed model (9125.84), and the range normalization model (9302.53).

Next, we examined the predictors of the divisive normalization model to determine whether participants exhibited violations of the IIA. As predicted by any model of value-based choice and replicating Study 1a, we found that the difference in value between target and competitor significantly predicted the likelihood that the target was hired, $b = 10.91$ 95% CI [9.56, 12.28], $p < .001$. In other words, as the value difference between the target and the competitor increased, participants were more likely to select the target. We also replicated the effect of the distractor value from Study 1a, though it was only marginally significant, $b = 0.55$ 95% CI [−0.02, 1.12], $p = .057$. Again, while participants' choices exhibited IIA violations, these violations occurred in the opposite direction predicted by the divisive normalization model; as the value of the distractor increased, participants were more likely to hire the target.

Simulations revealed we had 49.3% 95% CI [46.16, 52.45] power to detect the effect of distractor value ($b = 0.55$) on choice and 100% 95% CI [99.63, 100] power to detect the effect of difference in value between target and competitor ($b = 10.91$) on choice. Sensitivity analyses revealed we would have been able to detect a minimum effect size of 0.81 for distractor value (with power = 0.8, and $\alpha = 0.05$).

4.3. Study 1c

Likelihood of hiring ratings were highly reliable across repetitions (e.g., median within-participant correlation: $r = 0.94$) and across our sample ($r = 0.93$ 95% CI [0.92, 0.94], $t(1588) = 100.09$, $p < .001$). The hireability ratings were also good predictors of participant's choices in Phase 2. Across the population, participants selected the target on 68.6% of trials, the target on 28% of trials, and the distractor on only 3.4% of trials. For the main analysis, we excluded the trials ($n = 447$) in which the distractor candidate was chosen (leaving $n = 12{,}803$).

We employed the same analysis strategy as in Studies 1a and 1b. We found that this time the untransformed model had the lowest BIC (13,908.40), followed by the divisive normalization model (14,054.52), and the range normalization model (14,112.38).

As predicted by any model of value-based choice, we found that the difference in value between target and competitor significantly predicted the likelihood that the target was hired, $b = 0.63$ 95% CI [0.57, 0.68], $p < .001$. However, this time the distractor value did not predict the likelihood that the target was hired, $b = 0.007$ 95% CI [−0.04, 0.05], $p = .78$.

Simulations revealed we had 6% 95% CI [4.61, 7.66] power to detect the effect of distractor value ($b = 0.007$) on choice and 100% 95% CI [99.63, 100] power to detect the effect of difference in value between target and competitor ($b = 0.63$) on choice. Sensitivity analyses revealed we would have been able to detect a minimum effect size of 0.067 for distractor value (with power = 0.8, and $\alpha = 0.05$).

4.4. Study 1d

Likelihood of hiring ratings were highly reliable across repetitions (e.g., median within-participant correlation: $r = 0.91$) and across our sample ($r = 0.88$ 95% CI [0.88, 0.89], $t(4498) = 126.97$, $p < .001$). The hireability ratings were also good predictors of participant's choices in Phase 2. Across the population, participants selected the target on 65.1% of trials, the competitor on 30.1% of trials, and the distractor on only 4.8% of trials. For the main analysis, we excluded the trials ($n = 1814$) in which the distractor candidate was chosen (leaving $n = 35{,}686$).

Replicating Studies 1a and 1b we found that the divisive normalization model had the lowest BIC (40,842.90), followed by the untransformed model (40,862.93), and the range normalization model (41,206.79).

Examining violations of the IIA, we again replicated Studies 1a and 1b: difference in value between target and competitor significantly predicted the likelihood that the target was hired, $b = 9.13$ 95% CI [8.50, 9.76], $p < .001$, as did the distractor value, $b = 0.58$ 95% CI [0.30, 0.85], $p < .001$. Once again, when including the random effect of participant in the model, violations occurred in the opposite direction predicted by a divisive normalization account; as the value of the distractor increased, participants were more likely to hire the target.

Simulations revealed we had 98.3% 95% CI [97.29, 99.01] power to detect the effect of distractor value ($b = 0.58$) on choice and 100% 95% CI [99.63, 100] power to detect the effect of difference in value between target and competitor ($b = 9.13$) on choice. Sensitivity analyses revealed we would have been able to detect a minimum effect size of 0.38 for distractor value (with power = 0.8, and $\alpha = 0.05$).

5. Discussion

Three out of four studies established IIA violations in the domain of hiring. Specifically, in Studies 1a, 1b, and 1d, we found that divisive normalization value coding best captured empirical choice patterns, but that the nature of the violation was opposite of that predicted by the divisive normalization account: higher distractor values made participants *more,* not less likely to choose the highest-value target.

One difference between our analyses and others' was that we employed a multi-level model that included the random effect of participant (e.g., Louie et al., 2013 did not nest data within participant; Gluth et al., 2018 conducted a two-step analysis). When we modeled the data as though all of the data points were independent (i.e., excluded the random effect of participant) we replicated the results predicted by a divisive normalization account: specifically, as the distractor value increased, likelihood of hiring the target *decreased* (see supplemental materials for corresponding regression tables, for each model, for each study). Thus, our data demonstrate that aggregation may obscure or

even invert the pattern of subject-level results (Turner et al., 2018; see also Davis-Stober, Park, Brown, & Regenwetter, 2016; Heathcote, Brown, & Mewhort, 2000). As such, one significant contribution of the current work is to identify that in some cases divisive normalization results may be driven by a statistical artifact known as Simpson's paradox. Our results suggest that the population level findings (when data are aggregated) are driven by individual differences in how participants assign value to the distractors. Specifically, people who tend to assign lower values to *distractors* on average are less likely to violate the IIA whereas people who assign higher values are more likely to violate the IIA.

The generalizability of our results is limited, however, because we used a naïve sample of undergraduates for Studies 1a-1c, and a mix of undergraduates and community members for Study 1d. In the real-world, actual hiring decisions are made by experts who may be less subject to context effects (Ratneshwar, Shocker, & Stewart, 1987). Therefore, we next examined whether these IIA violations occur when decision-makers are highly motivated to make a beneficial choice: in this case, in U.S. congressional elections.

6. Study 2: value coding models and congressional election outcomes

Inspired by studies of distractor effects in political elections (Hedgcock, Rao, & Chen, 2009; Pan et al., 1995), we use a combination of archival electoral data and survey data to test whether normalization models can explain the outcomes of three-way congressional elections. Specifically, we documented the results of past three-way U.S. congressional elections and tested whether electoral outcomes are best characterized by a divisive normalization account. Based on existing findings that competence, inferred from faces, is a robust predictor of political preferences (Olivola & Todorov, 2010; Todorov, Mandisodza, Goren, & Hall, 2005) we decided to use candidates' facial competence as a proxy for their value. Note that this approach differs from previous investigations of context-dependence in voting behavior because we aim to explain population-level electoral outcomes (not individuals' voting behavior).

To address this question, we had participants who were unfamiliar with the candidates rate the candidates' faces on competence (as well as several other attributes, for which we control) and then used those ratings—untransformed, divisively normalized, and range normalized—to test which model best predicted electoral outcomes.

7. Methods

7.1. Participants and exclusions

We recruited participants to rate candidates' faces via Amazon's Mechanical Turk platform. We aimed for 50 ratings along 4 attributes—competence, familiarity, attractiveness, age—for 694 unique candidate faces. To avoid rater fatigue, we asked each participant to rate a subset of 30 randomly selected faces on all four attributes. Therefore, we recruited 1204 participants to generate candidate attribute ratings (593 female, 607 male, 2 declined to answer; M_{age} = 35.74 years, SD = 11.04).

7.2. Materials

We cataloged 254 three-way Senate and House of Representative races (and their outcomes; the equivalent number of trials in Experiments 1 and 2), which have taken place over the last 22 years (races span 2014–1994 for the Senate, and 2014–2012 for the House). We found 55 races for the Senate, and 199 races for the House. Elections were only included if all three candidates were officially listed on the ballot (i.e., not write-ins) and backed by an official political party. We also only used races where we could find photographs of all three candidates. We located (via Wikipedia and other online sources) professional photographs of all candidates' faces, standardized the image size and backgrounds (grey), and cropped the photographs so they include only the candidates' shoulders and face.

7.3. Procedure

We presented each participant with 30 candidate faces (randomly selected from the full set), and asked them to rate each face on competence, attractiveness, familiarity, and age. We asked participants to work as quickly as possible and rely on 'gut instincts' when responding. Importantly, we never told them that these were the faces of political candidates. Participants saw the same 30 faces for each of the 4 dimensions. Participants made competence ratings first, followed by attractiveness, familiarity, and age. Ratings for competence, attractiveness, and familiarity were made on a 0–100 slider-bar from very incompetent/unattractive/unfamiliar to very competent/attractive/familiar. Again, the number associated with their response did not appear on the screen. Participants' made age ratings on a slider-bar from 0 to 100, however, in this case they could see the number associated with the scale position. We randomized the order in which they saw candidates within each attribute. Finally, we included a recognition question with the competence ratings asking participants to check a box if they recognized the person in the photograph.

7.4. Analyses and exclusions

At the level of ratings, we excluded the 1% trials on which participants reported that they recognized the candidate's face. We then computed each candidate's "value" on each of the four attributes. We examined histograms for all four attributes for each of the candidates and as expected, found some of them to be non-normally distributed with varying amounts of skew and kurtosis. As a result, we used the median as our measure of central tendency for each candidate's value on each attribute.

At the level of races, we excluded races (n = 4) in which each candidate did not receive at least 40 ratings on each of the 4 dimensions. This resulted in 250 three-way races (Senate: 52, House: 198) with 682 unique candidates from across 44 states that we used in our final analyses.

To compare the models predicting electoral outcomes we fit logistic regressions in R 3.5.0 (R Core Team, 2018). Sensitivity analyses were conducted using Monte Carlo simulation via the *simr* package (version 1.0.4; Green & MacLeod, 2016). We report post-hoc power for all parameters. Since our effects of interest are how inferred competence affects race outcomes, we focused our sensitivity analyses on this parameter. For each parameter, we ran 1000 simulations.

8. Results

As a first step, we fit a linear regression to examine whether competence, familiarity, attractiveness, and age predicted the percentage of votes each candidate received. This analysis allowed us to determine which attributes most likely acted as inputs to voters' choices. Multiple regression results indicated that the four predictors explained 18.3% of the variance ($F(4, 745) = 41.75$, $p < .001$, R^2 of 0.183). Replicating previous studies, we found that competence ($\beta = 0.319$ 95% CI [0.68,1.28], $p < .001$) and age ($\beta = 0.198$ 95% CI [0.26, 0.69], $p < .001$) were significant predictors of vote share, but familiarity ($\beta = 0.018$ 95% CI [−0.16, 0.25], $p = .669$) and attractiveness ($\beta = 0.050$ 95% CI [−0.13, 0.32], $p = .401$) were not. Simulations revealed we had 100% 95% CI [99.63, 100] power to detect the effect of inferred competence on percentage of votes each candidate received. Sensitivity analyses revealed we would have been able to detect a minimum effect size of $b = 0.44$ or $\beta = 0.143$ for competence (with power = 0.8, and $\alpha = 0.05$).

8.1. Competence

Of the 250 races we cataloged, 22 of the races ended up including at least two candidates with identical inferred competence values, which made it impossible for us to classify the three targets as target, competitor, and distractor, respectively. Of the remaining 228 races, the candidate with the most inferred facial competence (*target*) won 119 of the races, the candidate with the second most inferred facial competence (*competitor*) won 73 of the races, and the candidate with the least inferred facial competence (*distractor*) won 36 of the races. We excluded the 36 races in which the distractor won from the model comparison analyses. We fit logistic regressions with the remaining races ($n = 192$) to test whether the most competent-looking candidate won as a function of the difference in facial competence between target and competitor and the facial competence of the distractor. In order to investigate whether electoral outcomes were best characterized by models that included untransformed vs. divisive normalization vs. range normalization values as inputs, we compared models using BIC and found that the untransformed model had the lowest BIC (249.90), followed by the divisive normalization model (252.73), and the range normalization model (252.98). However, the conventional standard set by Raftery (1995) considers a difference of BIC smaller than 2 between two models as "barely worth mentioning." As such, we report the results of all three models here.

We found that the competence difference between target and competitor was not a significant predictor of the race outcome for the untransformed values ($b = 0.07$ 95% CI [−0.003, 0.151], $p = .073$), a significant predictor for the divisively normalized values ($b = 14.28$ 95% CI [1.61, 28.60], $p = .037$), but not a significant predictor when we used the range-normalized values ($b_r = 0.57$ 95% CI [−0.64, 1.82], $p = .359$). We also found that the inferred competence of the distractor was a significant predictor of race outcome for the untransformed values ($b = -0.07$ 95% CI [−0.12, 0.03], $p = .002$), divisively normalized values ($b = -14.56$ 95% CI [−27.36, −2.77], $p = .020$), and range normalized values ($b = -0.16$ 95% CI [−0.25, 0.08], $p < .001$). (Please see supplemental materials for same analysis with age; in short, the findings indicated that inferred age of the distractor did not predict race outcome.)

Simulations revealed we had 66% 95% CI [63.38, 69.33] power to detect the effect of distractor competence on race outcome for the divisive normalization model. Sensitivity analyses revealed we would have been able to detect a minimum effect size of $b = -16.15$ for competence (with power = 0.8, and $\alpha = 0.05$).

In sum, while we find that both competence and age significantly predict vote share for past Congressional elections, we only find evidence of violations of the IIA along the competence dimension. One possible explanation of this finding is that age is a ratio scale with constant intervals between values and a meaningful zero point. In other words, if voters have a heuristic that older candidates are more qualified, no distractor is going to make them believe a 55 year old is older than a 60 year old. As such, attributes like age may be less subject to context effects relative to trait attributions.

9. General discussion

Across five studies, we demonstrated that violations of the IIA can occur in two consequential social contexts—hiring scenarios and actual political election outcomes—and demonstrated that these effects were best explained by a domain-general value coding mechanism. In Study 1a we found that divisive normalization value inputs best characterized empirical choice patterns in hypothetical hiring decisions. Furthermore, we found evidence for IIA violations such that as the value of the distractor increased, participants were more likely to choose the target over the competitor. In Study 1b, we replicated these findings and found that they were robust even when targets in the trinary choice set were presented sequentially, though the effect of the distractor value was marginal. In contrast, in Study 1c we found that untransformed value inputs best characterized empirical choice patterns, and we found no evidence of IIA violations. To arbitrate among the significant effects in 1a and 1b versus the null findings in 1c we tripled our sample size in Study 1d. Study 1d replicated both the model comparison and IIA violation results of Studies 1a and 1b.

Study 2 extended the results of Studies 1a-1d in a real-world outcome: real-world congressional race outcomes. Replicating previous studies, we found that inferences of competence and age from candidates' faces (controlling for attractiveness and familiarity) predicted outcomes of U.S congressional elections. Though the model comparison results failed to identify a clear "winner" with regard to model fit on the competence or age dimensions, we found that real past electoral outcomes comported with a divisive normalization account using inferred-competence as a proxy for value: as the inferred competence of a third candidate increased, the likelihood of the most competent looking candidate relative to the 2nd ranked candidate decreased. Participants did not exhibit an IIA violation along the dimension of candidates' age.

9.1. Positive versus negative relationship between distractor value and target choice

As we noted above, the results of Studies 1a, 1b, and 1d run counter to the prediction made by the divisive normalization account: that increasing the value of the distractor should decrease the likelihood of choosing the highest-valued target. Because our analysis strategy differed from previous studies—we conducted mixed effects models to account for the within-participant nested structure of the data—we also analyzed our data excluding the random effect of participant (see supplemental materials). When we treated each trial as an independent observation we observed the opposite effect of the predictor: higher distractor values decreased picking the target in Studies 1a and 1b. Thus our results highlight the possibility that some previous divisive normalization results are driven by a statistical artifact (i.e., Simpson's paradox). This holds even within this paper: examining the distractor effect on electoral outcomes at the population level yielded the predicted negative effect of distractor value on selected the highest "value" candidate. To provide a more thorough test of these value-coding accounts, future work should model the data both in the aggregate and accounting for subject-specific patterns.

While range normalization did poorly in our model comparisons across studies, one account suggests that range normalization and divisive normalization might be implicated in different parts of the decision process (Soltani et al., 2012). Range normalization may be the mechanism by which individual features of each option are represented while divisive normalization underlies the coding of the overall value associated with selecting each option. Our individual versus population-level result differences might offer some insight into another distinction between these two models. When we took the nested within-participant data structure into account, we observed a positive relationship between the distractor value and choice accuracy, which is predicted by the range normalization account (even though the option values were best fit by a divisive normalization model). When we treated each observation as independent, the option values and choice outcomes were both best characterized by the divisive normalization account.

Though these models were appropriate for testing whether people exhibit violations of the IIA in social choice, there are other models which offer a more complete account of context-dependence in decision-making. Specifically, future research should disentangle at what stage nonlinearity may be introduced into the decision-making process (e.g., at feature encoding versus expression on the response scale) by modeling all of the potential processing stages from input to decision output (Busemeyer et al., 2019).

9.2. Conclusion

The current findings add to a growing literature comparing value coding models underlying context-dependence in consequential social decision-making contexts. Understanding the underlying mechanism of how we represent the value of individuals in a choice set, and how that mechanism determines, in the case of the current studies, hiring and voting choices, may complement prejudice-reduction strategies to bring about more consistent social decision-making across any context in which candidates are jointly evaluated. Furthermore, our computational modeling approach allows for greater predictive precision when different models make similar qualitative predictions by capitalizing on divergent quantitative predictions. Furthermore, these results indicate that these violations may arise from the fundamental coding mechanism of value itself, rather than (or in addition to) higher-order processes such as inconsistent weighting of candidate attributes or motivated reasoning—the usual targets of intervention and diversity efforts in industry and the public sphere. In other words, even if evaluators were unencumbered by stereotypes or bias at the decision-making stage, their choices might remain inconsistent due to the influence of value normalization at evaluation. Thus, our framework give us greater purchase on understanding when and how specific alternatives exert their effects on social evaluation and choice in any context in which candidates are jointly evaluated.

Open practices

Complete materials, data, and data analysis code for all studies are available for download at the Open Science Framework (OSF): https://osf.io/4zwsd.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2019.103847.

References

Anderson, N. H. (1971). Integration theory and attitude change. *Psychological Review, 78*(3), 171.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bateson, M., Healy, S. D., & Hurly, T. A. (2003). Context–dependent foraging decisions in rufous hummingbirds. *Proceedings of the Royal Society of London B: Biological Sciences, 270*(1521), 1271–1276.

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

Bohnet, I., Van Geen, A., & Bazerman, M. (2015). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science, 62*, 1225–1234.

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in Cognitive Sciences, 23*(3), 251–263.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience, 13*(1), 51–62.

Chang, L. W., & Cikara, M. (2018). Social decoys: Leveraging choice architecture to alter social preferences. *Journal of Personality and Social Psychology, 115*(2), 206–223.

Cikara, M., & Fiske, S. T. (2011). Bounded empathy: Neural responses to outgroup targets' (mis)fortunes. *Journal of Cognitive Neuroscience, 23*, 3791–3803.

Davis-Stober, C. P., Park, S., Brown, N., & Regenwetter, M. (2016). Reported violations of rationality may be aggregation artifacts. *Proceedings of the National Academy of Sciences of the United States of America, 113*(33), E4761–E4763.

Fechner, G. T. (1860). *Elemente der psychophysik*.

Furl, N. (2016). Facial-attractiveness choices are predicted by divisive normalization. *Psychological Science, 27*, 1379–1387.

Gluth, S., Spektor, M. S., & Rieskamp, J. (2018). Value-based attentional capture affects multi-alternative decision making. *eLife, 7*, e39659.

Green, P., & MacLeod, C. J. (2016). simr: An R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*(2), 185–207.

Hedgcock, W., Rao, A. R., & Chen, H. (2009). Could Ralph Nader's entrance and exit have helped Al Gore? The impact of decoy dynamics on consumer choice. *Journal of Marketing Research, 46*(3), 330–343.

Herne, K. (1997). Decoy alternatives in policy choices: Asymmetric domination and compromise effects. *European Journal of Political Economics, 13*(3), 575–589.

Highhouse, S. (1996). Context-dependent selection: The effects of decoy and phantom job candidates. *Organizational Behavior and Human Decision Processes, 65*, 68–76.

Howes, A., Warren, P. A., Farmer, G., El-Deredy, W., & Lewis, R. L. (2016). Why contextual preference reversals maximize expected value. *Psychological Review, 123*, 368–391.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research, 9*, 90–98.

Hurly, T. A., & Oseen, M. D. (1999). Context-dependent, risk-sensitive foraging preferences in wild rufous hummingbirds. *Animal Behaviour, 58*, 59–66.

Johnson, S. K., Hekman, D. R., & Chan, E. T. (2016). If there's only one woman in your candidate pool, there's statistically no chance she'll be hired. *Harvard Business Review, April, 26*, 2016.

Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review, 100*, 109–128.

Khaw, M. W., Glimcher, P. W., & Louie, K. (2017). Normalized value coding explains dynamic adaptation in the human valuation process. *Proceedings of the National Academy of Sciences, 114*(48), 12696–12701.

Latty, T., & Beekman, M. (2011). Irrational decision-making in an amoeboid organism: Transitivity and context-dependent preferences. *Proceedings of the Royal Society B: Biological Sciences, 278*, 307–312.

Leung, M. D., & Koppman, S. (2018). Taking a pass: How proportional prejudice and decisions not to hire reproduce gender segregation. *American Journal of Sociology, 124*(3), 762–813.

Louie, K., Grattan, L. E., & Glimcher, P. W. (2011). Reward value-based gain control: Divisive normalization in parietal cortex. *Journal of Neuroscience, 31*, 10627–10639.

Louie, K., Khaw, M. W., & Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences, 110*, 6139–6144.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York: Wiley.

Mellers, B. A., & Birnbaum, M. H. (1983). Contextual effects in social judgment. *Journal of Experimental Social Psychology, 19*(2), 157–171.

Noguchi, T., & Stewart, N. (2018). Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological Review, 124*(4), 512–544.

Norton, M., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology, 87* (817–813).

Oakes, P., Haslam, S. A., & Turner, J. C. (1998). The role of prototypicality in group influence and cohesion: Contextual variation in the graded structure of social categories. In S. Worchel, J. F. Morales, D. Paez, & J.-C. Deschamps (Eds.). *Social identity: International perspectives* (pp. 75–92). London, England: Sage.

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior, 34*(2), 83–110.

Pan, Y., O'Curry, S., & Pitts, R. (1995). The attraction effect and political choice in two elections. *Journal of Consumer Psychology, 4*, 85–101.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*(6), 407.

Pettibone, J. C., & Wedell, D. H. (2000). Examining models of nondominated decoy effects across judgment and choice. *Organizational Behavior and Human Decision Processes, 81*(2), 300–328.

R Core Team (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 111–163.

Rangel, A., & Clithero, J. A. (2012). Value normalization in decision making: Theory and evidence. *Current Opinion in Neurobiology, 22*, 970–981.

Ratneshwar, S., Shocker, A. D., & Stewart, D. W. (1987). Toward understanding the attraction effect: The implications of product stimulus meaningfulness and familiarity. *Journal of Consumer Research, 13*, 520–533.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review, 108*, 370–392.

Sedikides, C., Ariely, D., & Olsen, N. (1999). Contextual and procedural determinants of partner selection: Of asymmetric dominance and prominence. *Social Cognition, 17*(2), 118–139.

Sen, A. K. (1971). Choice functions and revealed preference. *The Review of Economic Studies, 38*, 307–317.

Shafir, S., Waite, T. A., & Smith, B. H. (2002). Context-dependent violations of rational choice in honeybees (*Apis mellifera*) and gray jays (*Perisoreus canadensis*). *Behavioral Ecology and Sociobiology, 51*, 180–187.

Simonsohn, U., & Gino, F. (2013). Daily horizons: Evidence of narrow bracketing in judgment from 10 years of MBA admissions interviews. *Psychological Science, 24*, 219–224.

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research, 16*, 158–174.

Smith, R. H., Diener, E., & Wedell, D. H. (1989). Intrapersonal and social comparison determinants of happiness: A range-frequency analysis. *Journal of Personality and Social Psychology, 56*(3), 317.

Soltani, A., De Martino, B., & Camerer, C. (2012). A range-normalization model of context-dependent choice: A new model and evidence. *PLoS Computational Biology, 8*, e1002607.

Stevens, S. S. (1961). To honor Fechner and repeal his law. *Science, 133*(3446), 80–86.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*, 1623–1626.

Trope, Y., & Mackie, D. M. (1987). Sensitivity to alternatives in social hypothesis-testing. *Journal of Experimental Social Psychology, 23*, 445–459.

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences, 113*, 3102–3107.
Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2018). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review, 125*(3), 329–362.
Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76*, 31–48.
Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management Science, 39*, 1179–1189.
Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*, 550–592.
Wedell, D. H., & Parducci, A. (1988). The category effect in social judgment: Experimental ratings of happiness. *Journal of Personality and Social Psychology, 55*(3), 341.
Wyer, N. A., Sadler, M. S., & Judd, C. M. (2002). Contrast effects in stereotype formation and change: The role of comparative context. *Journal of Experimental Social Psychology, 38*, 443–458.