The Love of Large Numbers Revisited: A Coherence Model of the Popularity Bias[☆]

Daniel W. Heck[a,*], Lukas Seiling[b], Arndt Bröder[b]

[a] *Philipps-Universität Marburg*
[b] *University of Mannheim*

ARTICLE INFO

ABSTRACT

Preferences are often based on social information such as experiences and recommendations of other people. The reliance on social information is especially relevant in the case of online shopping, where buying decisions for products may often be based on online reviews by other customers. Recently, Powell, Yu, DeWolf, and Holyoak (2017, *Psychological Science, 28*, 1432-1442) showed that, when deciding between two products, people do not consider the number of product reviews in a statistically appropriate way as predicted by a Bayesian model but rather exhibit a bias for popular products (i.e., products with many reviews). In the present work, we propose a coherence model of the cognitive mechanism underlying this empirical phenomenon. The new model assumes that people strive for a coherent representation of the available information (i.e., the average review score and the number of reviews). To test this theoretical account, we reanalyzed the data of Powell and colleagues and ran an online study with 244 participants using a wider range of stimulus material than in the original study. Besides replicating the popularity bias, the study provided clear evidence for the predicted coherence effect, that is, decisions became more confident and faster when the available information about popularity and quality was congruent.

1. Introduction

A recurrent research topic in judgment and decision making concerns the question of how people integrate multiple pieces of information when stating preferences or making probabilistic judgments (Dawes, 1979; Tversky, 1969). Some scholars have focused on whether decision makers consider all available information in a compensatory way or whether they ignore less relevant cues (e.g., Einhorn, 1970; Payne, Bettman, & Johnson, 1993). For instance, Gigerenzer and Todd (1999) proposed Take-The-Best as a non-compensatory, lexicographic decision-making heuristic that only relies on the most valid available cue — a strategy that can be highly adaptive in specific ecological environments (Davis-Stober, Dana, & Budescu, 2010). Other research has focused on how information from different sources (e.g., social and nonsocial cues) is integrated to make a decision (Collins, Percy, Smith, & Kruschke, 2011; Schrah, Dalal, & Sniezek, 2006). For instance, in the advice-taking literature, scholars have investigated how people weigh their own opinion or experience with the advice or recommendations

by experts or other reference groups (e.g. Bonaccio & Dalal, 2006; Yaniv, 2004).

In daily life, the relevance of both social and nonsocial information in decision making becomes particularly evident in the domain of online shopping. Many shopping websites provide their customers with objective information about the available products (Ranganathan & Ganapathy, 2002). In addition, it has become common practice that customers can rate the quality of products by writing a review and providing a rating (e.g., assigning between one and five stars to a product; Chevalier & Mayzlin, 2006). Other web page visitors then see a summary of the distribution of reviews usually presented as the average review score and the number of reviews for a product. Given that a majority of customers take online reviews into account (e.g., Chevalier & Mayzlin, 2006; Kee, 2008), the natural question arises *how* visitors use this kind of social information when deciding which product to buy.

Besides the obvious relevance for applied areas such as marketing research, the domain of online reviews is ideally suited to investigate the cognitive mechanism underlying information integration. Recently,

Product DQ

⭐⭐⯪☆☆ (150)

2.5

Product GW

⭐⭐⭐⭐⭐ (25)

4.9

1 – Definitely purchase Product DQ	2	3	4	5	6 – Definitely purchase Product GW

Fig. 1. Example of the decision task presented in the current experiment.

Powell, Yu, DeWolf, and Holyoak (2017) investigated how people integrate information provided by different cues when making preference judgments based on online review scores. The two cues of interest were the product's average review score and the number of review scores. This type of stimulus material has the interesting property that it provides both statistical and social information about a product: From a statistical perspective, a larger number of reviews indicates that the corresponding average review score has a smaller sampling variance and thus provides a more precise estimate of a products quality. This follows directly from the law of large numbers for sample means (Wasserman, 2004, p. 76). From a social perspective, the number of reviews provides online shoppers with social information whether a product is popular, indicating how many other customers already decided to buy it (Chen, 2008). Hence, a large number of reviews can be interpreted as a cue signaling the higher popularity of the more-rated product (Powell et al., 2017). Such decision processes where "people will be doing what others are doing rather than using their information" (Banerjee, 1992; p. 797) has been termed "herd behavior." Given the well-known limitations of people's ability to rely on statistical information in many contexts (e.g., Kahneman & Tversky, 1982; Tversky & Kahneman, 1974), it is an important research question whether and how individuals take sample size into account when making decisions based on social information.

1.1. Bayesian Reasoning and a Heuristic Social-Inference Model

To test whether decision makers consider the number of online reviews as statistical information (i.e., more reviews imply more precise average review scores) or as a social information (i.e., more reviews indicate more popular products), Powell et al. (2017) conducted two online studies. Using a controlled experimental design, participants were presented with the average review score and the number of reviews of two fictitious products (e.g., two mobile phone cases) as shown in Figure 1 . Based only on this information, participants stated their relative preference for one versus the other product on a 6-point rating scale.

To compare observed judgments against a normative baseline, Powell et al. (2017) proposed a Bayesian model according to which the true quality of a product is inferred based on the number of reviews n and the average review score x (for a review of Bayesian models of decision making, see Slovic & Lichtenstein, 1971). The model makes the psychological assumption that people appropriately process statistical information, meaning that they assign more weight to the average review score x when the corresponding number of reviews n increases. To formalize this intuition underlying the law of large numbers (Wasserman, 2004), the Bayesian model treats the true quality of a product as an unknown parameter θ and assumes that judgments are based on the posterior distribution of θ given the average review score x and the number of reviews n:

$$P(\theta|x, n) = \frac{P(x, n|\theta)P(\theta)}{P(x, n)}. \tag{1}$$

Based on the central limit theorem (Wasserman, 2004), Powell et al.

(2017) assumed that the likelihood $P(x, n \mid \theta)$ of the observed average review score and the number of reviews can be described by a t-distribution with degrees of freedom determined by the number of reviews n. Hence, the Bayesian model assumes that average review scores are more reliable when they are based on many reviews. As a prior distribution $P(\theta)$ for the true quality of a product, Powell et al. (2017) drew samples from the empirical distribution of average review scores on a shopping website. Thereby, the Bayesian model assumes that the distribution of true values θ can be approximated by the distribution of reviews.

To illustrate a qualitative difference of using the number of reviews as statistical versus social information, consider choosing between two products: a mobile phone case with $n_1 = 150$ reviews and an average review score of $x_1 = 3.1$ and another phone case with only $n_2 = 25$ reviews but an average review score of $x_2 = 3.4$. According to the Bayesian model, the number of reviews provides statistical information because the average review score is more reliable when the underlying number of reviews is large. Since the empirical mean of review scores in the relevant product category (i.e., mobile phone cases) is approximately $\bar{x} = 3.7$ (Powell et al., 2017), the Bayesian model of statistical reasoning implies that the more-rated phone case is actually expected to be "significantly worse" than the less-rated one (because the lower average score of $x_1 = 3.1$ is based on a larger number of reviews). However, when treating the number of reviews as social information, the more-rated phone case might appear to be more popular. In turn, individuals might prefer the more-rated product despite its slightly worse average review score.

Powell et al. (2017) tested the parameter-free predictions of the Bayesian statistical model in two studies that differed with respect to the difference in the number of reviews between the more-rated and the less-rated product. In both studies, the number of reviews had a strong positive effect on preferences, meaning that participants generally preferred the more-rated product. People often preferred the more-rated product even when it had a lower average review score than the less-rated product, in which case the Bayesian model implies that there is a high probability that the more-rated product is actually the worse product. Given that the general tendency to prefer more popular products violates the statistical principles underlying the normative Bayesian model, Powell et al. (2017) labeled this empirical phenomenon as a *popularity bias*.[1]

To explain this bias in favor of more popular products and to describe information integration in the domain of product reviews, Powell et al. (2017) proposed a heuristic social-inference model that "predicts a bias toward selecting more-reviewed products that is independent of other factors" (p. 1433). This model assumes that people do not consider the number of reviews n in a statistically appropriate way to judge the precision of the corresponding average score x. Instead, they

---

[1] When believing that other people have access to more information about the domain of interest than oneself, it may be optimal to take the decisions of others into account (Banerjee, 1992). Thus, labeling this empirical phenomenon as a *bias* may be misleading in some scenarios. However, for consistency with the paper by Powell et al. (2017), we still use the term *popularity bias*.

interpret the number of reviews as a generally positive cue indicating popularity. More precisely, the heuristic social-inference model assumes that people's preference for one product over the other is based on the additive combination of the difference in the average scores x_1 and x_2 and the difference in the number of reviews n_1 and n_2. Hence, the model assumes that people generally prefer products with a higher average review score and with more reviews, and that these two factors contribute independently to the stated preference judgments. Using a logistic regression model for dichotomized preference ratings, Powell et al. (2017) showed that the heuristic social-inference model had a good model fit in absolute terms and performed much better than the normative Bayesian model.

Irrespective of its good fit to data, a major limitation of the heuristic social-inference model concerns its status as a descriptive model that lacks a clear psychological explanation of the underlying processes of information integration (Simon, 1992). In fact, the core idea that participants simply weigh multiple cues additively can easily be tested by fitting a logistic-regression model with the average review scores and the number of reviews as predictors. However, the heuristic social-inference model remains silent with respect to the underlying processes of how these two pieces of information are integrated to form an overall preference judgment. To overcome this limitation, we propose a novel, coherence-based account of the formation of preferences based on social information that makes explicit assumptions about the underlying cognitive processes of integrating the average review score and the number of reviews into an overall preference rating. Importantly, the coherence model allows to derive novel empirical predictions which can be tested in the domain of online reviews.

1.2. Coherence-Based Judgments

Coherence theories of information integration assume that people strive for a coherent representation of all available information based on which they arrive at an overall judgment (Holyoak & Simon, 1999; Thagard, 1989). Across many studies, it has been shown that the principle of explanatory coherence can explain judgments and decision making across a wide range of phenomena such as probabilistic inferences (Glöckner, Hilbig, & Jekel, 2014; Heck, Hilbig, & Moshagen, 2017), recognition-based decisions (Glöckner & Bröder, 2014; Heck & Erdfelder, 2017), information search (Jekel, Glöckner, & Bröder, 2018), preference construction (Simon & Spiller, 2016), and legal decision making (Simon, Snow, & Read, 2004). Based on the theoretical principle of coherence, parallel constraint satisfaction models (Glöckner & Betsch, 2008; Simon, Krawczyk, & Holyoak, 2004; Thagard & Verbeurgt, 1998) formalize the underlying cognitive process of integrating the available information as the maximization of the coherence in an artificial neural network. Whereas information that is available to the decision maker is often modeled by structural properties of the network, preference states are modeled by changes in the energy level of the neural network (Glöckner et al., 2014). Using this formal representation, parallel constraint satisfaction models have been shown to predict empirical choice probabilities, confidence ratings, search directions, and response times across various choice situations (Glöckner & Betsch, 2008; Heck & Erdfelder, 2017; Jekel et al., 2018; Scharf, Wiegelmann, & Bröder, 2019).

Given that coherence theory is a general theory of information processing, it is straightforward to derive specific predictions for the domain of online product reviews. Just as the heuristic social-inference model, we assume that people prefer popular products with many reviews (Banerjee, 1992). Moreover, the relative weight of the number of reviews and the average score determines whether the more-rated or the less-rated product is preferred (Glöckner & Betsch, 2008). However, instead of assuming an independent contribution of the average score and the number of reviews, we distinguish between paired comparisons of two products in which the available information facilitates a coherent cognitive representation and those cases in which it does not

(Betsch & Glöckner, 2010). In congruent trials, the more-rated product has a *higher* average score than the less-rated product, which provides coherent information in favor of the more-rated product. In incongruent cases, however, the more-rated product has a *lower* average score than the less-rated product, which provides incoherent information about the product preference. When comparing these two scenarios, coherence-based models such as parallel constraint satisfaction predict that congruent cases result in relatively fast and confident judgments in favor of the more-rated product because the larger amount of reviews and the higher average review score provide coherent information (Betsch & Glöckner, 2010; Heck & Erdfelder, 2017). In contrast, incongruent cases result in cognitive incoherence, which needs to be resolved in order to provide an overall preference judgment. Accordingly, responses are predicted to be less confident and relatively slow.

Overall, these predictions imply the presence of an interaction between the number of reviews and the average review score depending on the congruency of both factors. However, Powell et al. (2017) did not test whether the effects of the number of reviews and the average score were strictly additive or whether they interacted.[2] In fact, the assumption of their social-inference model that cue information is additively combined has a long tradition in judgment and decision research (Anderson, 1981; Dawes & Corrigan, 1974), and linear models are often seen as a good approximation to judgment with often little need to add configural or interaction terms (Brehmer, 1994; Karelaia & Hogarth, 2008). However, without a statistical test of the interaction, it is not clear whether the two cues (number of reviews and average review score) are in fact strictly additively combined as predicted by the social-inference model.

In the following, we contest the validity of the social-inference model by showing that the congruence of available information affects preference ratings and response times as predicted by coherence theory. First, we reanalyze the data by Powell et al. (2017) to test for an interaction effect between these two types of information. Second, we report a new study that aimed at replicating the popularity bias for an extended stimulus space. By including review-score differences up to ± 1.8 on the five-star review scale, we provide a stronger test of the proposed coherence model than the two original studies which were restricted to paired comparisons with relatively small review-score differences (i.e., $|x_1 - x_2| \leq 0.3$). Moreover, we also collected response times to test a unique prediction of the proposed coherence model, namely, that responses are slower when the available information is incongruent (i.e., when the more-rated product has a lower average review score than the less-rated product).

2. Reanalysis of Powell et al. (2017)

As a first empirical test of the coherence-based account of the formation of preferences based on social information, we reanalyzed the freely available data by Powell et al. (2017). Similar to the original analysis, we first modeled dichotomized preference ratings using a logistic-regression model. However, instead of using a fixed-effects model, we fitted a generalized linear mixed-effects model with random intercepts for participants to account for the repeated-measures design of the experiment and possible heterogeneity of preferences across participants (Bates, Mächler, Bolker, & Walker, 2015). Moreover, to facilitate the inclusion of the interaction effect in the coherence model, we used a different dependent variable in the logistic-regression model. Whereas Powell et al. (2017) analyzed preferences for the left versus

---

[2] In fact, the authors concluded that "participants treated cues about choice outcomes and prevalence as independent and additive factors, without assuming any subtler interaction" (p. 1441). However, we thank Derek Powell for clarifying that the "subtler interaction" in this statement was intended as a reference to the relatively complex Bayesian model of statistical reasoning and not as a reference to a literal interaction in the statistical sense.
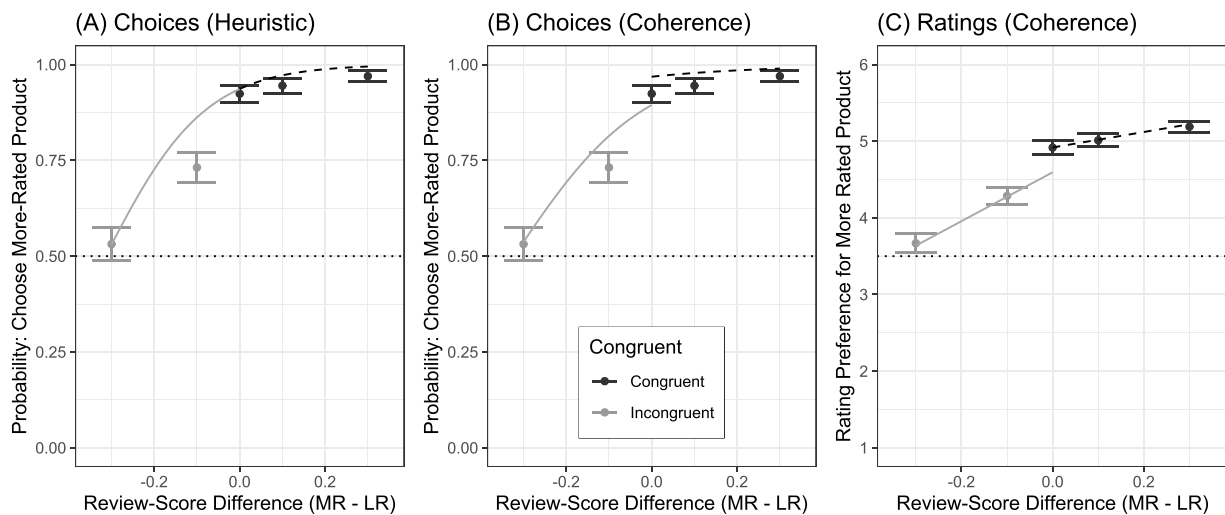
Fig. 2. Reanalysis of Study 2 by Powell et al. (2017). *Panel A:* The regression line shows the predicted probability of choosing the more-rated (MR) over the less-rated (LR) product based on the additive, heuristic social-cue model. *Panel B:* Choice probability fitted by the proposed coherence model. *Panel C:* Fit of the coherence model to observed preference ratings on a 6-point rating scale. Error bars show 95% confidence intervals.

the right product, we modeled preferences for the more-rated versus the less-rated product as the criterion. To test the heuristic social-inference model, we included the difference in the average review scores between the more-rated and less-rated product as a predictor. Moreover, the mean of the two average review scores was used as a control variable, thus allowing for possible effects of the location of both review scores on the five-star scale.

The fit of the mixed-effects logistic-regression model for the data of Study 2 by Powell et al. (2017) is shown in Figure 2 A. The fitted model curve shows that preferences for the more-rated product increase monotonically as the difference in the average review scores increases. Moreover, the popularity bias is reflected by an estimated intercept larger than 50% for the case that both products had the same average review score (i.e., $x_1 - x_2 = 0$), thus implying a general shift of preferences in favor of the more-rated product. Since the heuristic social-inference model assumes a simple additive effect of the review-score difference and the popularity effect (i.e., an upwards shift of the intercept), the fitted regression line is a smooth function irrespective whether the more-rated product has a higher or lower average score than the less-rated product.

To test the coherence-based account, we extended the logistic-regression model by adding another predictor, namely, a dummy-coded congruency variable with values of 0 and 1 encoding whether the average score of the more-rated product was either *strictly smaller* or *larger or equal* to that of the less-rated product, respectively. By including both the main effect of this congruency variable and its interaction with the difference of average scores, the logistic regression provides a test of the coherence model for preference ratings. Figure 2 B shows that this model assumes two separate regression lines for congruent (black) and incongruent (gray) trials, meaning that the effect of the difference of the average review scores depends on the coherence of the available information. As predicted by the coherence model, congruent information lead to a strong positive preference for the more-rated and better-rated product, whereas incongruent information resulted in less extreme preference ratings. Moreover, the impact of the review-score difference varied between the two cases as shown by the different slopes. Note that the regression curve for incongruent choices (gray) did not perfectly fit the average choice frequencies due to the inclusion of random intercepts in the nonlinear logistic link function. Nevertheless, Figure 2 shows that the more complex coherence model in Panel B had a better fit to the observed means than the heuristic social-inference model with simple additive effects in Panel A, an impression that was supported by a nested likelihood-ratio test, $\chi^2(2)$

$= 35.7, p < .001$.

Originally, Powell et al. (2017) dichotomized the observed 6-point preference ratings as being smaller or larger than 3.5 to facilitate the comparison of the data with the (binary) Bayesian model. However, to base our analysis on more fine-grained information, we also analyzed the actually observed preference ratings using a mixed-effects linear regression model with random intercepts and the same predictor variables as for the logistic regression. Figure 2 C shows that the coherence effect emerged even more clearly for these more granular data as shown by the different intercept and slope of congruent versus incongruent ratings. In line with this visual impression, the coherence model again resulted in a significantly better fit than the additive heuristic model ($\chi^2(2) = 104.8, p < .001$). To quantify effect size, we computed the increase in explained variance ΔR^2 for the fixed-effects terms of the linear mixed-effects models (Johnson, 2014; Nakagawa & Schielzeth, 2013). Whereas the explained variance of the two fixed-effects terms in the additive, social-inference model was $\Delta R^2 = .177$, the inclusion of congruency and its interaction with the review-score difference by the coherence model resulted in an increase of explained variance of $\Delta R^2 = .022$.[3]

Similar to the results for Study 2 of Powell et al. (2017), a reanalysis of the data of their Study 1 also supported the coherence model irrespective whether the logistic choice model ($\chi^2(2) = 49.9, p < .001$) or the linear rating model was used ($\chi^2(2) = 193.6, p < .001$). Again, the fixed-effects terms of the additive linear model explained a substantial amount of variance in ratings ($\Delta R^2 = .200$), with a moderate increase of $\Delta R^2 = .032$ for the coherence model. The evidence for an interaction of the predictors with respect to preferences clearly speaks against the additive effect of popularity and average review score as formulated in the original social-inference model. Rather, review-score differences had a stronger impact on preferences in incongruent than in congruent cases.

Overall, our results support the coherence-based account of information integration. However, the reanalyses are limited by the small number of data points and the restricted range of review-score differences. Figure 2 shows that the experimental design included only five levels of review-score differences and that the coherence model provides a lot of flexibility to account for the corresponding observed

---

[3] We are reporting the *marginal* R^2 measure proposed by Nakagawa and Schielzeth (2013), which focuses on the explained variance of the fixed-effects terms only. If the random-effects terms are considered as well, the coherence model resulted in a *conditional* explained variance of $R^2 = .468$

means. Hence, in the following study, we implemented a larger range of differences in the average review scores and elicited response times to provide a stronger test of the coherence model.

3. Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

3.1. Participants

To replicate the results by Powell et al. (2017), we aimed for a sample size larger than that of the original two studies (i.e., $N = 105$ and $N = 132$, respectively). We chose a minimum sample size of $N = 200$, which ensures that the standard error of the relative frequency of preferring the more-rated product was smaller than 0.035 for each of the paired comparisons. We provided participants with the opportunity to complete the study in either German or English. This was facilitated by the stimulus material, which was not language-specific (cf. Figure 1), and provided us with the opportunity to recruit both German participants from the University of Mannheim and English-speaking participants via the social media platform Reddit (Shatz, 2017). As a compensation, university students received course credit, and every participant had the chance to win one out of ten online-shopping vouchers each with a value of 10 €. The total number of recruited participants was $N = 254$. After exclusion of participants (see below), the sample size for analysis was $N = 244$ (mean age = 25.0, SD = 9.1; 68.4% female) including 185 German- and 59 English-speaking participants.

3.2. Materials and Design

The study implemented a within-subjects design in which each participant rated the relative preference for two fictitious products in 75 paired comparisons. For each of the two products, the average review score was displayed visually by the corresponding proportion of filled stars out of five total stars with the exact value displayed directly below (cf. Figure 1). Moreover, the number of reviews was shown in parentheses next to the stars for each product. In each trial, the two products differed in popularity with the more-rated product having approximately 150 reviews and the less-rated product having approximately 25 reviews. Following Powell et al., a repeated presentation of these two numbers was avoided by randomly drawing the exact number of reviews of the more-rated product from a uniform distribution between 145 and 155 while ensuring a constant difference of 125 relative to the less-rated product (Powell et al., 2017). The position of the more-rated product was counterbalanced for each participant so that both the more-rated and the less-rated product appeared equally often on the left or right side of the screen.

The average review scores of both the less-rated and the more-rated product were varied over nine levels between 4.9 and 2.5 in steps of 0.3. This range of review scores was centered on 3.7 which is close to the empirical mean of average review scores (Powell et al., 2017). By combining all nine levels of the average review scores for the more-rated and the less-rated product, we obtained a 9×9 grid of paired comparisons with review-score differences ranging from -2.4 to 2.4 in steps of 0.3. Out of these 81 paired comparisons, the order of 75 trials with review-score differences between $-1.8, -1.5, -1.2, …, +1.8$ was randomized and used for the analysis below. The six items with the most extreme review-score differences (± 2.1 and ± 2.4) were shown as filler trials at fixed positions (i.e., in trial 1, 14, 27, etc.) to increase trial heterogeneity and thus maintain a high level of involvement. Note that the filler item with the most extreme advantage for the more-rated product (with a review-score difference of $+2.4$) was used as an attention check item assuming that participants did not follow instructions or responded randomly if they preferred the less-rated product

with an average review score of 2.5 over the more-rated product with an average score of 4.9.

3.3. Procedure

The experimental setup aimed at replicating the conditions by Powell et al. (2017) as closely as possible and was programmed and made available via SoSci Survey (Leiner, 2018). At the start of the experiment, participants gave their informed consent and were informed about the procedure, the estimated time for completion, the possible compensation, and confidentiality. Adapting the task description of Powell et al. (2017), participants were informed that in each trial, they had to decide between two mobile phone cases based on the summary information for a set of reviews from a popular online shopping website. There was no further description apart from the fact that both phone cases were similarly priced. The instructions were shown for at least eight seconds to ensure that participants read all necessary information.

Next, participants were presented with the 81 paired comparisons including experimental trials in random order and filler and check items at fixed positions. Each of the products were labeled by two uppercase letters such as "FX" versus "ZA", and no product label appeared twice during the experiment. As a response scale, we used a 6-point rating scale with a rating of 1 representing a strong preference for the product shown on the left, and a rating of 6 representing a strong preference for the product on the right. We also recorded response time as the time interval between the initial presentation of the choice condition and the click on one of the six rating-scale buttons (cf. Figure 1). Directly after clicking on the button, the choice was logged to the data base; thus, participants did not have the possibility to change their response.

At the end of the study, participants answered two demographic questions (age and gender). Moreover, as a proxy for online buying experience, we asked how often participants made online purchases with the response options "less than once a year" ($N = 15$), "once a year" ($N = 47$), "once a month" ($N = 155$), "once a week" ($N = 23$), and "more than once a week" ($N = 4$). We also administered a seriousness-check question asking whether participants had completed the study seriously or just clicked through (Aust, Diedenhofen, Ullrich, & Musch, 2013). On the final pages, participants were thanked for their participation and could enter their email address to sign up for the lottery and receive an email with information about the research goal and results of the study.

4. Results

4.1. Exclusion of participants and trials

Of the 254 participants completing the study, six were removed from the analysis due to giving a negative response to the seriousness-check question. Moreover, in the attention-check trial, four participants indicated a preference in favor of the less-reviewed product with an average review score of 2.5 over the more-rated product with a score of 4.9. These four participants were removed under the assumption of not paying attention to the task at hand, thus resulting in a sample size of $N = 244$ for the analysis.

We also filtered trials with response-time outliers. Separately for each participant, we removed trials with response times that were outside three times the interquartile range (Tukey, 1977). Moreover, we removed trials with extreme response times faster than 300ms or slower than 20,000ms. The filtering of extreme response times let to the exclusion of 536 out of 18,300 trials.

4.2. Replication of Powell et al. (2017)

To facilitate a comparison with the predictions of the Bayesian model, Powell et al. (2017) dichotomized the observed preference
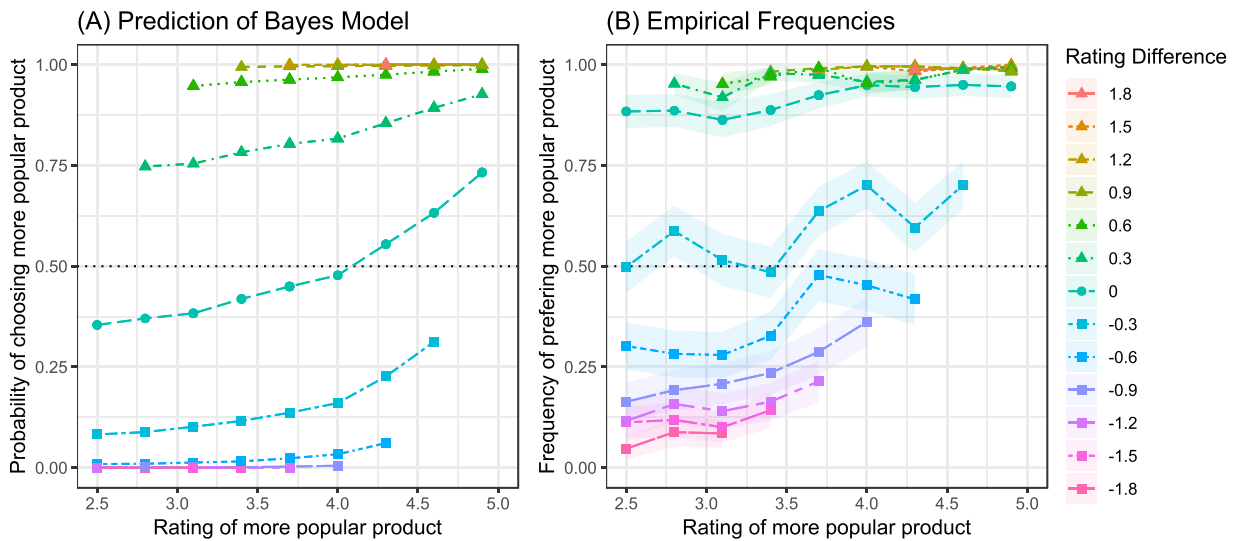
Fig. 3. *Panel A:* Probability of choosing the more-rated product as predicted by the parameter-free Bayesian model for the extended stimulus set used in the present study. Note that the lines overlap for absolute review-score differences larger than 0.9. *Panel B:* Relative frequency of choosing the more-rated product based on dichotomized preference ratings. The corresponding ribbons show 95% confidence intervals for the relative frequencies.
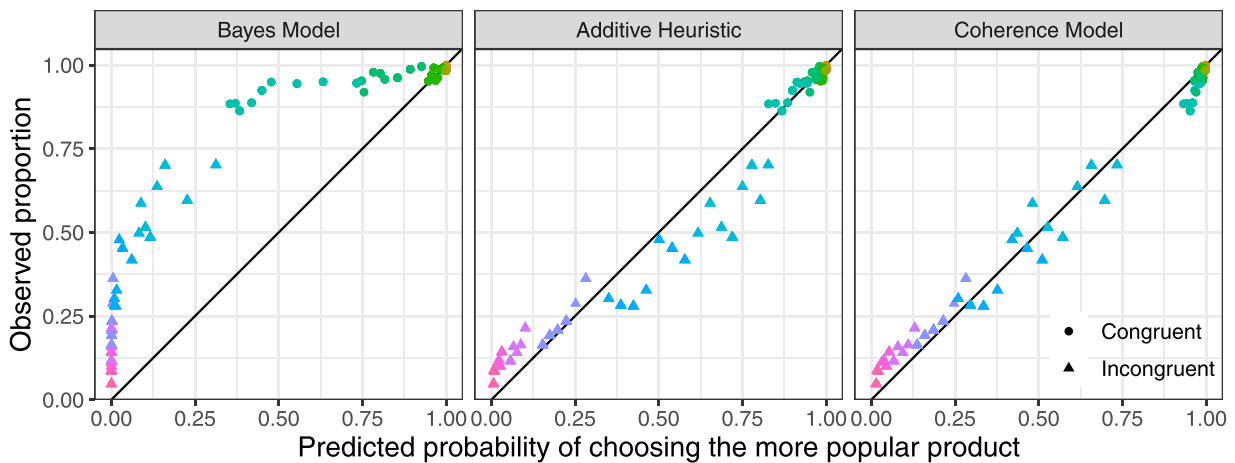
Fig. 4. Model fit of the Bayesian model, the additive social-inference heuristic, and the coherence model. The color key refers to review-score differences between the two presented products (cf. Figure 3).

ratings into binary choices, with values of 1 representing choices of the more-rated product and 0 representing choices of the less-rated product. Figure 3 shows that the observed choice frequencies made by the participants (Panel B) differed qualitatively from the predictions of the Bayesian Model (Panel A) as simulated for the extended stimulus set. The discrepancy is most obvious for paired comparisons with a review-score difference of zero (indicated by a dashed line with rounded points): Whereas the Bayesian model predicted relatively ambiguous choices with probabilities around 50%, empirical choice frequencies for the more-rated product were larger than 80% irrespective of the average review score of the more-rated product (as shown on the x-axis). Similarly, empirical frequencies for the other review-score differences (indicated by separate lines) were generally shifted upwards in favor of the more-rated product relative to the predictions of the Bayesian model, thus replicating the popularity bias. The degree of misfit becomes even more evident when plotting the predicted probabilities of choosing the more-rated product against the empirically observed proportions as shown in the first panel of Figure 4. For almost all paired comparisons, the more-rated product was preferred to a much higher degree than predicted by the Bayesian model.

The conclusions of these graphical comparisons of model fit were corroborated by sign tests addressing whether more people chose the

more-reviewed product in each of the paired comparisons (using a significance level of $\alpha = .01$). To replicate the analysis of Powell et al. (2017), we first restricted the tests to those 25 conditions with review-score differences of -0.3, 0.0, and $+0.3$ which were also included in the original studies. For these paired comparisons, sign tests showed that participants preferred the more-rated product in 22 of the 25 cases. When excluding cases in which the Bayesian model was nearly indifferent (i.e., when the predicted choice probability was between $.45 < P < .55$), the model favored the *less-reviewed* product in 13 of 24 cases. However, for these 13 paired comparisons in which the Bayesian model predicted choosing the less-reviewed product, sign tests indicated a preference for the *more-rated* product in 10 cases (and were not significant otherwise), meaning that participants often did not favor the product predicted by the Bayesian model. Descriptively, participants preferred the more-reviewed product in 70.5% of these paired comparisons, which is similar to the observed percentage of 65.5% reported by Powell et al. (2017).

Next, we compared the predictions of the Bayesian model for the set of all 75 paired comparisons with review-score differences between -1.8 and $+1.8$. Sign tests with a significance level of $\alpha = .01$ showed that participants preferred the more-rated product in 45 out of the 75 paired comparisons. When excluding cases in which the Bayesian model

was nearly indifferent, the model predicted a preference for the more-rated product in 36 of 74 conditions. In all of these paired comparisons, participants did indeed favor the more-reviewed product. However, in the remaining 38 paired comparisons for which the Bayesian model predicted a preference for the less-rated product, sign tests showed that participants preferred the less-rated and the more-rated product in 10 and 22 cases, respectively. Across these 38 trials for which the Bayesian model predicted a preference for the less-reviewed option, participants preferred the more-reviewed product in 38.5% of the trials.

These analyses show that the Bayesian model performed slightly better when including paired comparisons with larger review-score differences between -1.8 and $+1.8$ compared to using a restricted stimulus set with smaller differences (Powell et al., 2017). A similar tendency can also be observed in Figure 3 A showing that empirical frequencies were closer to the predictions of the Bayesian model for larger absolute review-score differences compared to the original conditions with relatively small review-score differences of -0.3, 0.0, and $+0.3$. Overall, the results show that the popularity bias is a robust empirical phenomenon that is pronounced most strongly if the difference in the average review scores is relatively small.

4.3. Testing the Coherence Model

To test the proposed coherence model, we fitted separate regression models to predict the three different dependent variables (i.e., choices and preference ratings for the more-rated product and response times).[4] Similar to the reanalysis above, we used logistic regression to predict choices of the more-rated product (i.e., dichotomized preference ratings; Powell et al., 2017) and linear regression to predict 6-point preference ratings. Moreover, we used a linear regression to predict log response times. Each of the three models included the same four predictor variables as in the reanalysis in Section 2: First, the difference in the average review scores of the more-rated and the less-rated product; second, the congruency dummy variable with a value of 0 if the more-rated product had a lower average review score and 1 otherwise; third, the interaction of review-score difference and congruency; and forth, the mean review score of the two products which served as a control variable. In all models, we included a random intercept to account for possible differences between participants. All generalized linear mixed models were fitted in R using the package lme4 (Bates et al., 2015). Data and R scripts are available on the Open Science Framework (https://osf.io/mzb7n/).

For each of the three dependent variables, we tested the coherence model against the additive, heuristic social-inference model which does not include the main effect of congruency and its interaction with the review-score difference as predictors. The corresponding likelihood-ratio tests indicated a significant discrepancy between the two models for all three dependent variables including the logistic regression of dichotomized choices ($\chi^2(2) = 453.6$, $p < .001$), the linear regression of preference ratings ($\chi^2(2) = 3376.1$, $p < .001$), and the linear regression of log response times ($\chi^2(2) = 582.2$, $p < .001$). With respect to predictive performance, the fixed-effects terms of the heuristic model explained $\Delta R^2 = .544$ of the variance in preference ratings, with the coherence model having an incremental validity of $\Delta R^2 = .064$. For log response times, the additive model explained a smaller proportion of observed variance ($\Delta R^2 = .052$), with the coherence model again having substantial incremental validity ($\Delta R^2 = .020$).[5]

In addition to these quantitative measures of model fit, Figure 4 provides a direct comparison of how well the Bayesian model, the heuristic social-inference model, and the coherence model predicted the observed choice proportions. Whereas the Bayesian model generally underestimated the popularity of the more-rated products, the additive heuristic model showed a tendency to overestimate the popularity of more-rated products that had only a slightly worse average review score than their less-rated competitor (i.e., incongruent cases with review-score differences of -0.3 and -0.6). Note that these paired comparisons are the ones in which the incoherence of the available information is maximal, and in turn, these cases were better described by the coherence model.

Overall, these results provide evidence in favor of the coherence model which assumes a main effect and interaction of congruency. The corresponding parameter estimates, standard errors, and p-values for the regression coefficients are reported in Table 1 . Moreover, for the linear regression models of ratings and log response times, Table 1 also shows the proportion of unique variance ΔR^2 explained by each of the fixed-effects terms of the coherence model.

Figure 5 shows that the coherence model had a very good fit for each of the three dependent variables. For dichotomized choices (Panel A), relative frequencies of preferring the more-rated product were close to one, in line with the prediction that coherent information results in stronger preference for the better-rated product with more reviews. In contrast, choice frequencies were around 50% when the conflict between the available information was large, that is, when the more-rated product was slightly worse than the less-rated product. Once the difference in average scores became larger, preferences for the better-rated product with less reviews became stronger. The same pattern emerged when analyzing observed preference ratings instead of dichotomized responses (Panel B). In this case, the discrepancy between incongruent and congruent conditions became even more pronounced as indicated by the larger gap of the regression lines for a review-score difference of zero. Most importantly, we also found a coherence effect for response times. As predicted, responses were on average faster in congruent than in incongruent conditions. Moreover, in both cases, responses were slowest when the conflict between the available information was largest (i.e., when review-score differences were close to zero). Once the difference in the average review scores increased, response times became faster both for congruent and incongruent conditions. These results are in line with the coherence-based account according to which decision conflict decreases when the review-score difference becomes very large (in which case it is easier to achieve a coherent representation since the number of reviews has less weight; Heck & Erdfelder, 2017).

5. Discussion

By replicating the main findings of Powell et al. (2017), we showed that the popularity bias is an empirically robust phenomenon, meaning that people generally tend to prefer more-rated over less-rated products. In contrast to the predictions of a Bayesian model of statistical reasoning (Powell et al., 2017), this also holds if the more-rated product has a lower average review score than the less-rated product. Going beyond a direct replication, we showed that the popularity bias also holds for a larger range of review-score differences, although the effect is strongest when the average review scores of two products are similar.

To explain the cognitive mechanisms underlying preference judgments in the domain of online reviews, we proposed a coherence-based account of preference formation based on the integration of social information. Assuming that people strive for a coherent representation of

---

[4] Given that "dichotomized choices" are obtained by a deterministic transformation of the observed preferences ratings, the two corresponding analyses are statistically dependent. Nevertheless, we decided to report the logistic regression to facilitate a comparison with the original analyses by Powell et al. (2017).

[5] When considering the conditional R^2, which also considers the random intercepts (Nakagawa & Schielzeth, 2013), the proportion of variance in ratings

(footnote continued)

and log response times explained by the coherence model was $R^2 = .694$ and $R^2 = .412$, respectively.

Table 1
Estimated regression coefficients for the coherence model.

	Estimate	SE	df	t	p	ΔR²
Choices: Hierarchical logistic regression						
(Intercept)	−0.979	0.213		−4.59	< .001	
Review-score difference (MR–LR)	2.943	0.090		32.88	< .001	
Congruency (0 = no; 1 = yes)	2.096	0.106		19.73	< .001	
Difference × Congruency	−0.995	0.180		−5.54	< .001	
Mean review score	0.606	0.048		12.74	< .001	
Preference ratings: Hierarchical linear regression						
(Intercept)	2.896	0.056	2640	51.46	< .001	
Review-score difference (MR–LR)	1.142	0.021	17521	55.65	< .001	.053
Congruency (0 = no; 1 = yes)	1.043	0.025	17520	41.53	< .001	.013
Difference × Congruency	−0.750	0.026	17520	−29.14	< .001	.030
Mean review score	0.316	0.012	17519	27.20	< .001	.015
log Response times: Hierarchical linear regression						
(Intercept)	8.895	0.036	1136	248.30	< .001	
Review-score difference (MR–LR)	0.123	0.011	17518	10.71	< .001	.004
Congruency (0 = no; 1 = yes)	−0.184	0.014	17517	−13.11	< .001	.016
Difference × Congruency	−0.339	0.014	17518	−23.60	< .001	.006
Mean review score	−0.144	0.006	17517	−22.28	< .001	.018

Note. The review-score difference refers to the difference of average review scores between the more-rated (MR) and the less-rated (LR) product. For the logistic regression, the *t*-value column contains normally-distributed *z*-statistics (hence, the degrees of freedom are missing). For the two linear models, the effect size ΔR² shows the proportion of unique variance explained by each of the regression terms.

all available information (Simon, Pham, Le, & Holyoak, 2001; Thagard & Verbeurgt, 1998), we predicted a coherence effect for preference ratings and response times: In addition to the simple main effects of average score and number of reviews, the speed and confidence of preference judgments should depend on the match between the two cues, that is, whether the more-rated product also has a higher average score. Thereby, the model differs from the heuristic social-inference model proposed by Powell et al. (2017), which assumes simple additive effects of the number of reviews and the average review score on preference ratings and remains silent with respect to response times.

We tested the proposed coherence model in a reanalysis of the data

by Powell et al. (2017) as well as in a new empirical study in which we extended the range of review-score differences presented to the participants. Using generalized mixed effects models, we defined a dummy-coded predictor to model congruent and incongruent cases in which the more-rated product had the higher or lower average review score, respectively. The analyses showed that both preference ratings and response times were strongly affected by this congruency variable and its interaction with the review-score difference, thus corroborating the proposed coherence model. As predicted, preference ratings were more extreme and faster when the available information was congruent, that is, when the more-rated product had a higher average review score. In contrast, preference ratings were more ambiguous and slower when the available cues provided conflicting information, that is, when the more-rated product had a lower average score. This effect emerged most clearly for highly incoherent cases, that is, for paired comparisons in which the more-rated product had only a slightly worse average review score. These cases produced very large response times (see Figure 5) and choice proportions that were fitted much better by the coherence model than by the social heuristic model (see Figure 4), thus corroborating the proposed coherence-based account of information integration.

5.1. Coherence-Based Accounts in Judgment and Decision Making

The proposed coherence model provides a psychological explanation of how social information is integrated into the formation of preferences. According to the theory of coherence as a general principle of information integration (Thagard, 1989), preferences and holistic judgments are formed by searching for a coherent internal representation of the available information. If a choice option is favored on all dimensions, the available information is congruent and the model predicts that people respond faster and with higher confidence (Glöckner & Betsch, 2008). In contrast, if different choice options are favored on different dimensions, the available information is incongruent and the model predicts that people respond more slowly and with lower confidence. By showing that these systematic patterns of preference ratings and response times also emerge in preference formation based on social information (i.e., product reviews by other customers), our work extends the scope of coherence as a general theory of integrating different types of information into a holistic judgment (Glöckner et al., 2014).
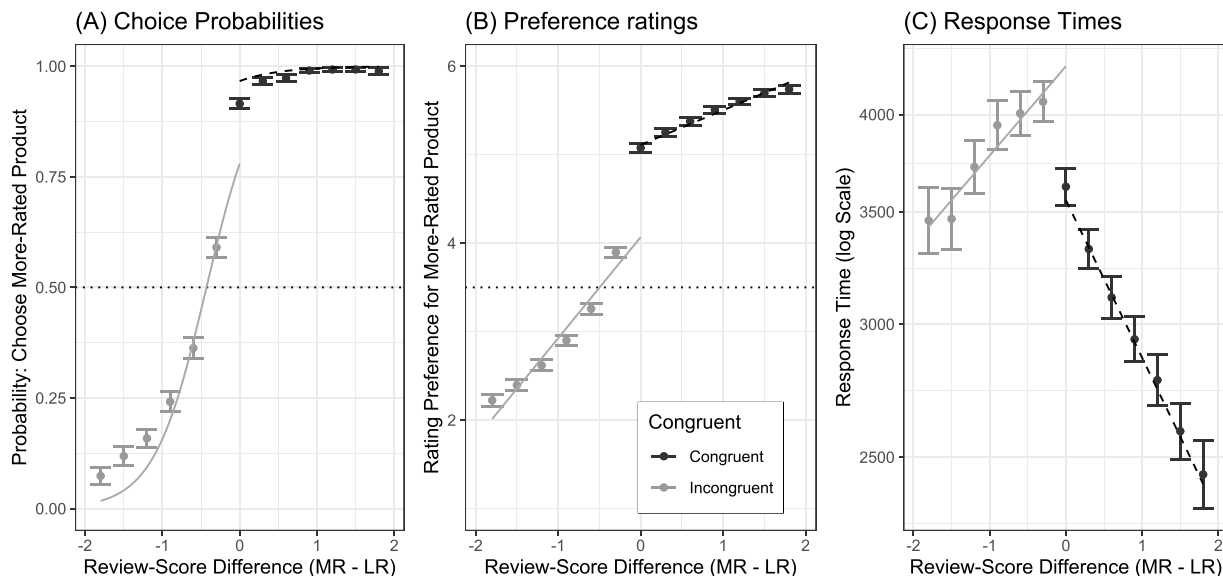
Fig. 5. *Panel A:* Relative frequency of choosing the more-rated product. The regression lines show the predicted probabilities of the proposed coherence model. *Panel B:* Fit of the coherence model to observed preference ratings. *Panel C:* Fit of the coherence model to log response times. Error bars show 95% confidence intervals.

Our results are in line with a growing literature providing evidence for the validity of coherence-based accounts for many phenomena and contexts in judgment and decision making. Originally, explanatory coherence was proposed as a theory of how people reason in everyday life when faced with a set of explanatory hypotheses (Thagard, 1989), for instance, when forming impressions of other individuals based on stereotypes, traits, and behavior (Kunda & Thagard, 1996). Extending these ideas to preference judgments in multiattribute decision making, Simon, Krawczyk, et al. (2004) showed that preferences are dynamically constructed by maximizing the coherence of the available information. The theory successfully predicted that preferences develop during the time course of decision making (Simon et al., 2001), that more information is processed faster than less (Betsch & Glöckner, 2010), and that, after a decision, the relative weight of cues is adjusted in line with the given choice (Glöckner, Betsch, & Schindler, 2010; Simon, Krawczyk, et al., 2004). More recently, computational models of coherence for probabilistic inferences have been developed to explain how people search for missing information in multiattribute decisions (Jekel et al., 2018; Scharf et al., 2019) and how they integrate information from memory during information integration (i.e., whether a choice option is recognized; Glöckner & Bröder, 2014; Heck & Erdfelder, 2017). The present work builds on these findings, showing that a coherence-based account can also explain how social information is used in preference formation.

To explain how people rely on social information, we adapted an interpretation of coherence-based theories assuming "fast, automatic processes that lead to consistent mental representations of the task and intuitive choices that emerge without awareness of the process itself" (p. 642; Glöckner et al., 2014). According to this view, the polarity of the available cues is jointly evaluated by fast, dynamic processes as modeled by parallel constraint satisfaction theory, a network model that simulates the activation of a set of bidirectionally connected nodes (Glöckner & Betsch, 2008; Read, Vanman, & Miller, 1997). Importantly, our definition of coherence thus differs from an alternative interpretation according to which people are assumed to adhere to *logical coherence* as required by normative theories of rational decision making (Arkes, Gigerenzer, & Hertwig, 2016; Hammond, 2000). In fact, such a normative view served as the basis for Powell et al.'s Bayesian model which treats the number of reviews as statistical information and could not account for the popularity effect found in the observed preference ratings.

In the present context of product preferences, one might think of yet another interpretation of coherence in terms of the *plausibility* of a distribution of product reviews with prior knowledge (Connell & Keane, 2006).[6] When asked to give a preference judgment for the two products, participants might construct a mental model of other customers (Johnson-Laird, 1980), assuming that people usually prefer products with high average review scores. According to such a mental model, it is implausible (i.e., "incoherent") that a product with many reviews has a low average review score because people would have stopped buying the product, implying that a large number of reviews could not have accumulated.

However, such a mental model of other customers would be at odds with the actual distribution of reviews on shopping websites, given the absence of an ecological correlation between the number of reviews and the average review score (Powell et al., 2017). Moreover, it is difficult to reconcile how the plausibility account could explain the pattern of observed response times shown in Figure 5 . For incongruent cases, the mental model becomes more implausible as the absolute review-score difference in favor of the less-rated product increases. When assuming that the construction and evaluation of mental models requires more time with decreasing plausibility, it follows that response times should increase. However, Figure 5 clearly shows that mean response times

decreased for larger absolute review-score differences in incongruent cases, thus providing evidence against the hypothesis that participants assessed the plausibility of a mental model. In contrast, the coherence model proposed in the present paper assumes a fast, automatic process of information integration and thus predicts this decrease in response times. Essentially, incoherence of the available information is maximal if the more-rated product has only a slightly worse average review score, whereas incoherence decreases if the more-rated product is clearly worse than the less-rated product. Note that a similar effect has been found in multiattribute decision making, where adding positive cues to a slightly favored option leads to faster responses even though more information has to be processed (Betsch & Glöckner, 2010; Heck & Erdfelder, 2017). Finally, given that our experimental paradigm lacked details about the specific products or context, we think that the task fostered fast, automatic responses as opposed to elaborate, effortful processes as required for the construction and evaluation of mental models.

5.2. Limitations and Future Directions

In the controlled experimental setting proposed by Powell et al. (2017), participants had to state preferences for fictitious mobile phone cases based only on the average review score and the number of reviews. This design has the advantage that one can directly draw conclusions about the impact of the number of reviews on preferences without having to account for actual differences in product quality as required in correlational studies (e.g., Chevalier & Mayzlin, 2006). However, despite this advantage of having a high internal validity, the controlled experimental design cannot be used to test the relative impact of the number of online reviews compared to other information such as longer text reviews in which customer's describe their experiences with a product (Sridhar & Srinivasan, 2012). Still, the core finding that people show a preference for more popular products has also been supported by studies in marketing research, which focus on the prediction of buying decisions based on customer reviews. In several of these studies, larger numbers of reviews were associated with increased sales ranks of books (e.g., Chevalier & Mayzlin, 2006; Sun, 2012) or video games (Zhu & Zhang, 2010). However, in contrast to these correlational analyses of sales statistics, the experimental setting used by Powell et al. (2017) allows to test causal effects of the average score and the number of reviews irrespective of differences in the true quality of products.

Besides the number of reviews and the average score, the present study ignored other features of the distribution of online reviews that may affect preferences. For instance, Sun (2012) showed that the standard deviation of the distribution of product reviews also affects sales ranks, a finding that is in line with the proposed coherence-based account. Given that a large variability of products indicates incoherence of the available information, preference construction should require more time and result in more ambiguous preferences compared to cases where reviews are more homogeneous. Similarly, empirical distributions of online reviews are often bimodal, meaning that most reviews are either very positive or very negative (Hu, Pavlou, & Zhang, 2006), an effect possibly due to the self-selection of reviewers on shopping websites (Hu, Zhang, & Pavlou, 2009). Again, a high degree of bimodality indicates a larger degree of incoherence of the available information and should result in similar patterns of preference ratings as high standard deviations. However, since studies in marketing research have usually focused on the prediction of sales ranks (Hu et al., 2006; Sun, 2012), it might be necessary to adapt a more controlled experimental design such as that used in the present study to disentangle the effect of other features of the distribution of reviews.

Going beyond the domain of online reviews, future research could elaborate and test novel predictions of the proposed coherence model for the integration of social information. First, one could investigate whether the direction of incoherence matters. In fact, it could be more

---

[6] We thank Derek Powell for bringing up this alternative explanation.

incoherent to have many reviews for a poorly rated product than few reviews for a highly-rated product. However, since participants only provided a single relative judgment for both products in our study, we cannot test for a possible asymmetry between the incoherence of a less-rated product with a high score and a high-rated product with a low score. Second, the paradigm could be extended in order to test the prediction of a post-decisional coherence shift, that is, whether the relative weight of the number of reviews increases after deciding in favor of the more-rated product (e.g., Simon, Krawczyk, et al., 2004). Finally, future research could focus on developing a computational model of the proposed coherence-based account, for instance, by adapting previous implementations of parallel constraint satisfaction theory for multiattribute decision making (Glöckner & Betsch, 2008; Glöckner et al., 2014).

5.3. Conclusion

The domain of online reviews provides a rich source of information for testing theories about preference construction based on social information. First, in times of the Internet, online reviews are of a high practical relevance for customer decisions and marketing. Second, the number of reviews has a special theoretical property, namely, that the sample size provides both statistical information (i.e., about the reliability of the average score) and social information (i.e., about the popularity of a product). Besides replicating the empirical finding that people prefer popular products with many reviews (Powell et al., 2017), we showed that the popularity bias also holds for a wider range of stimulus materials than that used in the original study. Moreover, we proposed a coherence explanation of empirical patterns in the observed preference ratings and response times. Thereby, the present study provides evidence that a coherence-based theory of information integration can account for preference formation based on social information.

References

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2016). How bad is incoherence? *Decision, 3*, 20–39 doi:10.1037/dec0000043.

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods, 45*, 527–535 doi:10.3758/s13428-012-0265-2.

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics, 107*, 797–817 doi:10.2307/2118364.

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

Betsch, T., & Glöckner, A. (2010). Intuition in judgment and decision making: Extensive thinking without effort. *Psychological Inquiry, 21*, 279–294 doi:10.1080/1047840X.2010.517737.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101*, 127–151 doi:10.1016/j.obhdp.2006.07.001.

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica, 87*, 137–154 doi:10.1016/0001-6918(94)90048-5.

Chen, Y.-F. (2008). Herd behavior in purchasing books online. *Computers in Human Behavior, 24*, 1977–1992 doi:10.1016/j.chb.2007.08.004.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*, 345–354 doi:10.1509/jmkr.43.3.345.

Collins, E. C., Percy, E. J., Smith, E. R., & Kruschke, J. K. (2011). Integrating advice and experience: Learning and decision making with social and nonsocial cues. *Journal of Personality and Social Psychology, 100*, 967–982 doi:10.1037/a0022982.

Connell, L., & Keane, M. T. (2006). A model of plausibility. *Cognitive Science, 30*, 95–120 doi:10.1207/s15516709cog0000_53.

Davis-Stober, C. P., Dana, J., & Budescu, D. V. (2010). Why recognition is rational: Optimality results on single-variable decision rules. *Judgment and Decision Making, 5*, 216–229.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American psychologist, 34*, 571–582 doi:10.1037/0003-066X.34.7.571.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106 doi:10.1037/h0037613.

Einhorn, H. J. (1970). The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin, 73*, 221–230 doi:10.1037/h0028695.

Gigerenzer, G., Todd, P. M. the ABC Research Group (1999). Simple heuristics that make

us smart. New York, NY: Oxford University Press.

Glöckner, A., & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making, 3*, 215–228 Retrieved from http://journal.sjdm.org/bn3/bn3.html.

Glöckner, A., Betsch, T., & Schindler, N. (2010). Coherence shifts in probabilistic inference tasks. *Journal of Behavioral Decision Making, 23*, 439–462 doi:10.1002/bdm.668.

Glöckner, A., & Bröder, A. (2014). Cognitive integration of recognition information and additional cues in memory-based decisions. *Judgment and Decision Making, 9*, 35–50 Retrieved from http://journal.sjdm.org/13/13912/jdm13912.html.

Glöckner, A., Hilbig, B. E., & Jekel, M. (2014). What is adaptive about adaptive decision making? *A parallel constraint satisfaction account, Cognition, 133*, 641–666 doi:10.1016/j.cognition.2014.08.017.

Hammond, K. R. (2000). Coherence and correspondence theories in judgment and decision making. In T. Conolly, H. R. Arkes, & K. R. Hammond (Eds.). *Judgment and decision making: An interdisciplinary reader* (pp. 53–65). .

Heck, D. W., & Erdfelder, E. (2017). Linking process and measurement models of recognition-based decisions. *Psychological Review, 124*, 442–471 doi:10.1037/rev0000063.

Heck, D. W., Hilbig, B. E., & Moshagen, M. (2017). From information processing to decisions: Formalizing and comparing probabilistic choice models. *Cognitive Psychology, 96*, 26–40 doi:10.1016/j.cogpsych.2017.05.003.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General, 128*, 3–31 doi:10.1037/0096-3445.128.1.3.

Hu, N., Pavlou, P. A., & Zhang, J. (2006). Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication. *In Proceedings of the 7th ACM Conference on Electronic Commerce* (pp. 324–330).

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped Distribution of Product Reviews. *Communications of the ACM, 52*, 144–147 doi:10.1145/1562764.1562800.

Jekel, M., Glöckner, A., & Bröder, A. (2018). A new and unique prediction for cue-search in a parallel-constraint satisfaction network model: The attraction search effect. *Psychological Review, 125*, 744–768 doi:10.1037/rev0000107.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science, 4*, 71–115 doi:10.1016/S0364-0213(81)80005-5.

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution, 5*, 944–946 doi:10.1111/2041-210X.12225. pmid: 25810896.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition, 11*, 123–141 doi:10.1016/0010-0277(82)90022-1.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*, 404–426 doi:10.1037/0033-2909.134.3.404.

Kee, T. (2008). *Majority of online shoppers check at least four reviews before buying.* Retrieved from https://www.mediapost.com/publications/article/76727/majority-of-online-shoppers-check-at-least-four-re.html.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*, 284–308.

Leiner, D. J. (2018). *SoSci Survey.* Retrieved from https://www.soscisurvey.com.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*, 133–142 doi:10.1111/j.2041-210x.2012.00261.x.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* Cambridge University Press.

Powell, D., Yu, J., DeWolf, M., & Holyoak, K. J. (2017). The love of large numbers: A popularity bias in consumer choice. *Psychological Science, 28*, 1432–1442 doi:10.1177/0956797617711291.

Ranganathan, C., & Ganapathy, S. (2002). Key dimensions of business-to-consumer web sites. *Information & Management, 39*, 457–465 doi:10.1016/S0378-7206(01)00112-4.

Read, S. J., Vanman, E. J., & Miller, L. C. (1997). Connectionism, parallel constraint satisfaction processes, and Gestalt principles: (Re)introducing cognitive dynamics to social psychology. *Personality and Social Psychology Review, 1*, 26–53 doi:10.1207/s15327957pspr0101_3.

Scharf, S. E., Wiegelmann, M., & Bröder, A. (2019). Information search in everyday decisions: The generalizability of the attraction search effect. *Judgment and Decision Making, 14*, 488–512 Retrieved from http://journal.sjdm.org/18/181123/jdm181123.pdf.

Schrah, G. E., Dalal, R. S., & Sniezek, J. A. (2006). No decision-maker is an island: Integrating expert advice with information acquisition. *Journal of Behavioral Decision Making, 19*, 43–60 doi:10.1002/bdm.514.

Shatz, I. (2017). Fast, free, and targeted: Reddit as a source for recruiting participants online. *Social Science Computer Review, 35*, 537–549 doi:10.1177/0894439316650163.

Simon, D., Krawczyk, D. C., & Holyoak, K. J. (2004a]). Construction of preferences by constraint satisfaction. *Psychological Science, 15*, 331–336 doi:10.1111/j. 0956-7976.2004.00678.x. pmid: 15102143.

Simon, D., Pham, L. B., Le, Q. A., & Holyoak, K. J. (2001). The emergence of coherence over the course of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1250–1260 doi:10.1037/0278-7393.27.5.1250.

Simon, D., Snow, C. J., & Read, S. J. (2004b]). The redux of cognitive consistency theories: Evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology, 86*, 814–837 doi:10.1037/0022-3514.86.6.814.

Simon, D., & Spiller, S. A. (2016). The elasticity of preferences. *Psychological Science, 27*,

1588–1599 doi:10.1177/0956797616666501.

Simon, H. A. (1992). What is an "explanation" of behavior? Psychological Science, 3, 150-161. doi:10.1111%2Fj.1467-9280.1992.tb00017.x.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6*, 649–744 doi:10.1016/0030-5073(71)90033-X.

Sridhar, S., & Srinivasan, R. (2012). Social influence effects in online product ratings. *Journal of Marketing, 76*, 70–88 doi:10.1509/jm.10.0377.

Sun, M. (2012). How does the variance of product ratings matter? *Management Science, 58*, 696–707 Retrieved from https://www.jstor.org/stable/41432790.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435–467 doi:10.1017/S0140525X00057046.

Thagard, P., & Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science, 22*, 1–24 doi:10.1207/s15516709cog2201_1.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76*, 31–48 doi:10.1037/h0026750.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131 doi:10.1126/science.185.4157.1124.

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference.* New York, NY: Springer.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93*, 1–13 doi:10.1016/j.obhdp.2003.08.002.

Zhu, F., & Zhang, X. M. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing, 74*, 133–148 doi:10.1509/jm.74.2.133.