Journal of Behavioral Decision Making, *J. Behav. Dec. Making*, **28**: 214–223 (2015)
Published online 26 September 2014 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/bdm.1841

Online Product Reviews and the Description–Experience Gap

DIRK U. WULFF[1]*, THOMAS T. HILLS[2] and RALPH HERTWIG[1]
[1]*Max Planck Institute for Human Development, Berlin, Germany*
[2]*University of Warwick, Coventry, UK*

ABSTRACT

People can access information about choices in at least two ways: via summary descriptions that provide an overview of potential outcomes and their likelihood of occurrence or via sequential presentation of outcomes. Provided with the former, people make *decisions from description*; with the latter, they make *decisions from experience*. Recent investigations involving risky choices have demonstrated a robust and systematic description–experience gap. Specifically, when people make decisions from experience, rare events tend to have less impact than what they deserve according to their objective probability. Here, we show that this description–experience gap generalizes from choices involving monetary gambles to choices based on (hypothetical) online product ratings. We further show that causes that have been identified in the context of risky choice also contribute to the description–experience gap in choice based on online product ratings: reliance on relatively small samples of information and overweighting of recently sampled information (recency). We conclude with a discussion of the practical implications of our results and identify promising directions for cross-disciplinary investigations. Copyright © 2014 John Wiley & Sons, Ltd.

KEY WORDS word of mouth; online consumer ratings; Amazon; description–experience gap; information search; decisions under uncertainty

Neoclassical theory of consumer behavior (e.g., Marshallian demand) conceives consumer choice as choice under certainty. Challenging this conception, Savage (1954/1972) emphasized the importance of uncertainty in decisions about consumer products:

> Jones is faced with the decision whether to buy a certain sedan for a thousand dollars, a certain convertible also for a thousand dollars, or to buy neither and continue carless. The simplest analysis, and the one generally assumed, is that Jones is deciding between three definite and sure enjoyments, that of the sedan, the convertible, or the thousand dollars. Chance and uncertainty are considered to have nothing to do with the situation. […] however, it is not difficult to recognize that Jones must in fact take account of many uncertain future possibilities in actually making his choice. (pp. 83–84)

One source of uncertainty—and the one that is the concern of the present article—is the degree to which the consumer will be satisfied with his choice. Is driving a convertible really as fun as Jones expected it to be? How much will he enjoy driving it in the winter? How worried should he be about sun exposure? Fortunately, individual consumers are not alone when faced with these uncertainties. With the rise of the Internet and social media, more than ever before, consumers can learn from the experience of others. Indeed, online product reviews provide a specific form of vicarious experience that has become ubiquitous. In the fast-growing market of electronic business-to-consumer commerce (U.S. Census Bureau, 2009), they have become a market force in their own right, successfully mediating online purchasing activity (e.g., Dellarocas, 2003).

Numerous investigations have demonstrated how product reviews and ratings can affect book sales (Chevalier & Mayzlin, 2006) and box office revenues (Duan, Gu, & Whinston, 2008; Liu, 2006) or boost growth in preferences for certain types of beers (Clemons, Gao, & Hitt, 2006). To the best of our knowledge, most studies examining the link between product reviews and sales figures have analyzed large-scale panel data (e.g., Chen, Wu, & Yoon, 2004; Chevalier & Mayzlin, 2006; Dellarocas, Zhang, & Awad, 2007; Duan et al., 2008). Thus, previous research on online product reviews has predominantly taken the seller's perspective. The consumer perspective and the question of how consumers process online product ratings have received less attention. This study helps to fill this gap by taking advantage of recent findings from behavioral decision making and demonstrating how they pertain to online product reviews. It also contributes to the growing literature on online decision making (Darley, Blankson, & Luethge, 2010; Punj, 2012) that addresses, in particular, the uncertainty associated with the lack of direct experience with a product or with sales staff (Johnson, Bellman, & Lohse, 2003), as well as the information search required prior to making a selection (Horrigan, 2008; Peterson & Merino, 2003).

Parallels between risky choice and online product reviews

Although choice between consumer products is not identical with choice between monetary gambles, there are similarities between the two: A single online consumer rating can be conceived as a potential future state of satisfaction after the purchase of the product. Thus, when a consumer seeks to buy a particular product, she may assume that her future satisfaction equals the satisfaction of the person who previously purchased the product and provided the rating. If there are many similar ratings of the product, she can assume that she will be as happy as all previous buyers.

*Correspondence to: Dirk U. Wulff, Max Planck Institute for Human Development, Center for Adaptive Rationality (ARC), Lentzeallee 94, 14195 Berlin, Germany. E-mail: wulff@mpib-berlin.mpg.de

However, if variance occurs among raters, she will be uncertain as to which of the potential satisfaction levels will apply to her. Under the simplifying assumption that she has no additional information, she will have to assume that each individual rating in the full set of ratings (one by each rater) has the same likelihood of matching her future satisfaction level. It follows that the set of consumer ratings for a product, when aggregated by rating categories, can be understood as a gamble over states of satisfaction, where the relative frequencies of rating categories indicate the probability of future states of satisfaction.

This investigation seeks to use the resemblance between these two choice situations to create a bridge between the two fields of research. To this end, we provide an example of how the literature on risky choice can inform research on online consumer choice. Specifically, we capitalize on two dimensions that play an important role in both online product reviews and recent investigations of risky and uncertain choice: format of information presentation and distributional characteristics.

Electronic commerce (e-commerce) sites like Amazon.com display the overall mean of the available consumer ratings as a number of stars. In addition, they present a summary bar plot and a list of individual ratings. Formally, both formats present identical distributional information, but they differ substantially in the way users experience that information. Summary bar plots present complete information in one descriptive format. Individual ratings, in contrast, require the user to *sequentially search* through the ratings to acquire representative information. The distinction between summary bar plots and individual ratings can be mapped onto a distinction between two formats of information representation that has received much attention in recent investigations of risky choice involving monetary gambles. The distinction, detailed in the succeeding texts, is that between *decisions from experience* and *decisions from description* (Hertwig, Barron, Weber, & Erev, 2004). Numerous studies have demonstrated that these two kinds of formats and decisions can result in systematic and predictable differences in choices, the *description–experience gap* (for reviews, refer to Hertwig & Erev, 2009; Rakow & Newell, 2010).

The second parallel between research on risky choice and online product reviews is the *bimodal* nature of the outcome distribution. Risky choice is often studied using two-outcome gambles (Holt & Laury, 2002; Kahneman & Tversky, 1979). In many cases, these two-outcome gambles comprise a probable outcome and a complementary (relatively) rare event (Erev et al., 2010). Analyzing ratings from Amazon.com, Hu, Zhang, and Pavlou (2009) found that most distributions of online product ratings follow a J-shaped[1] pattern: many very positive ratings, few very negative ratings, and hardly any ratings in between. Hu et al. (2009) suggested two selection biases to explain this distribution. First, people who give a product a low valuation are less likely to purchase it and therefore less likely to submit a rating relative to customers who actually purchased the product. Furthermore, among the purchasers, those who arrive at an extreme—either positive or negative—valuation of a product are more likely to express their views than are those with less extreme valuations, leading to a bimodal distribution (with the positive mode being more frequent than the negative one). The resulting J-shaped distribution can be conceived of as an extension of a two-outcome risky gamble containing a rare event.

In what follows, we briefly introduce relevant findings from recent research on the description–experience gap. We then explore how these findings can be brought to bear on consumer choices, based on "experienced" and "described" product reviews.

The description–experience gap

In most studies of risky choice, people are provided with a summary description of the risky options. The options' outcomes and associated probabilities are either conveyed visually (e.g., by a pie chart or frequency distribution) or described using numbers in text. An example of a summary description is as follows:

Option A: Receive \$4 with probability of .8, \$0 otherwise.

or

Option B: Receive \$3 for sure.

When outcomes and chances are presented in this *description* format, the majority of people choose option B (e.g., Hertwig et al., 2004; Kahneman & Tversky, 1979), even though option A has the higher expected value (A, \$3.2 vs B, \$3). This phenomenon has often been explained as a consequence of the propensity to overweight rare events; that is, people choose as if they overweight the small probability of winning nothing in gamble A (Kahneman & Tversky, 1979).

Another way to learn about the outcomes and their likelihoods is to experience those outcomes iteratively over a series of samples. For example, an onlooker witnessing the outcomes sampled from options A and B may see the following distribution of associated payoff schedules:

Option A: \$0, \$4, \$4, \$0, \$4, \$4, and \$4

Option B: \$3, \$3, \$3, \$3, \$3, \$3, and \$3

In the laboratory version of this *sampling paradigm*, participants can experience as many outcomes as they wish without the associated monetary consequences, before then deciding to terminate the exploration period and make a final choice. When gamble outcomes are presented in this *experience* format, people predominantly choose option A (Hertwig et al., 2004; Ungemach, Chater, & Stewart, 2009; but refer to Hills, Noguchi, & Gibbert, 2013). This reversal of preference implies that when decisions are based on

[1]The term "J-shaped" has two possible meanings: Sometimes, it is used to refer to a unimodal power-law distribution (e.g., Anderson & Schooler, 1991; Hertwig, Hoffrage, & Sparr, 2012; Todd & Gigerenzer, 2007), in which few objects have extreme values and most objects have small to medium values; sometimes, it is used to refer to a bimodal distribution (refer to Vokó et al., 1999; Witteman et al., 1994). The latter meaning is the one used here.

DOI: 10.1002/bdm

experience, people choose as if rare events received less weight than what they deserve in light of their objective probabilities. The description–experience gap in choice has been replicated across a wide range of studies (for reviews, refer to Hertwig & Erev, 2009; Rakow & Newell, 2010).

Why are rare events underweighted in experienced-based choices? Several reasons have been proposed (Hertwig & Erev, 2009). The two most important ones that pertain to online reviews are limited search and recency. Time constraints limit a person's ability to explore infinitely. Furthermore, there is evidence that people may be content with only small amounts of information, as small samples amplify the difference between options, thus easing choice difficulty (Hertwig & Pleskac, 2010). However, small samples also bear the risk that the decision maker is not informed about the existence of rare events or that the rare event is represented less often than expected (refer to Hertwig et al., 2004).

Another, though less powerful, factor is *recency* (compare, e.g., Hertwig et al., 2004; Rakow, Demes, & Newell, 2008; Ungemach et al., 2009). Outcomes occurring later in the sampling sequence seem to have more impact than earlier samples (Hertwig et al., 2004). This could be caused by memory limitations (e.g., Murdock, 1962) or be the outcome of an information updating process (Hogarth & Einhorn, 1992). As a consequence of recency, a decision maker who performed sampling extensively may nevertheless rely on a functionally small sample. When the functional sample size is constrained to recent samples, rare events are unlikely to be incorporated in the person's final assessment of the option.

Does the description–experience gap generalize to choices based on online product reviews?

In summary, the situation in which people make product choices based on online product reviews has much in common with the various formats in which risky decision making has been studied. First and most importantly, in both choice situations, people make choices over probability distributions of outcomes—monetary rewards in studies on risky choice and levels of satisfaction in online consumer choice. Second, the distributions of outcomes in both situations are bimodal, with one mode being rare—usually the extreme negative mode in online product ratings. Third, the formats of information presentation used either display summary presentations of the outcomes (ratings) or require self-paced sequential search.

Despite these parallels between online consumer choice and risky choice, the two research fields have remained largely unconnected. We explore one possible link by testing whether behavioral effects documented for abstract monetary gambles generalize to choices between consumer products based on consumer ratings. The potential synergies for both domains are promising. To summarize, the rich experimental and theoretical literature on risky choice can serve as a starting point to overcome the lack of experimental work on individual consumer choice. Online consumer choice, in turn, represents an increasingly germane real-world choice scenario that can be used to test the generality of the effects found with monetary gambles. The description–experience gap has often been demonstrated using two-outcome gambles, rendering this investigation an extension not only in terms of the type of outcome but also in terms of the payoff distributions' complexity.

Does a description–experience gap also exist in choices based on online product ratings? To answer this question, we conducted a laboratory experiment in which participants chose between two products (e.g., camcorders) solely on the basis of product ratings. We varied the presentation of these ratings between a full summary (description format) and one requiring participants to search through a series of individual ratings (experience format). We examined the extent to which these description-based versus experience-based formats triggered systematically different choice proportions, mirroring those found in investigations of risky choice. In other words, we examined whether the experience format, relative to the description format, resulted in people choosing as if they underweighted rare (extreme) ratings relative to their objective probabilities. Moreover, we examined the extent to which two cognitive mechanisms observed as contributing to the description–experience gap in risky choice, *limited search* and *recency*, also operate in choice based on consumer ratings. Specifically, we predicted that avid searchers are more likely to have experienced rare product ratings than frugal searchers and are therefore less likely to choose as if they underweight rare ratings. In accordance with Hertwig and Pleskac's (2010) finding, we further predicted that frugal searchers will judge their decision to be easier than avid searchers, irrespective of the information experienced. Finally, we predicted that ratings experienced later in the sampling sequence will have more impact than those experienced earlier (recency effect).

METHOD

Participants

We collected data from 63 participants (43 female participants). The mean age was 27 years. Participants were rewarded by either course credit or a fixed payment of Confoederatio Helvetica franc (CHF) 15 (~$15.00) and also received a monetary bonus based on the outcomes of their choices. Specifically, a random draw was taken out of each chosen rating distribution, and the resulting value of the rating (the number of stars) was multiplied by CHF 0.05. On average, participants earned a bonus of CHF 3.56 (~$3.5).

Procedure and material

Participants made 10 hypothetical choices between pairs of consumer products, once in the description format and once in the experience format. For each choice, product images were presented next to each other on the computer screen. We collected product images from several e-commerce sites to cover a wide range of applications and price ranges (e.g., laptops, restaurant dinners, pairs of shoes, coffee makers, etc.). The respective consumer ratings were displayed below each product (either in a summary plot or as individual ratings). Apart from consumer ratings, no further information

DOI: 10.1002/bdm

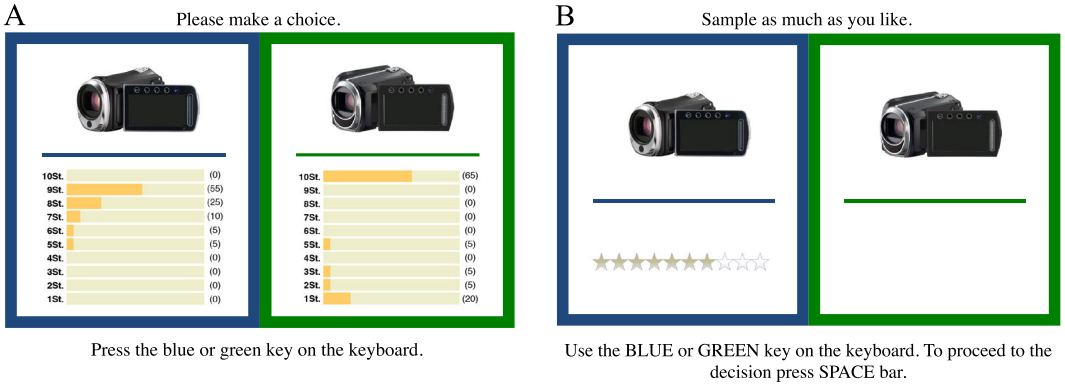

Figure 1. Screenshot of the description (A) and experience (B) rating formats. In the description format, a full table of 100 ratings is displayed, one below each product. In the experience format, ratings are presented individually below each product as it is sampled. The occurrences of individual ratings are determined by the underlying distributions (identical to those in the description format) and the search behavior of the participant.

was provided. Participants were instructed to select the product that appealed most to them given the distribution of ratings. To encourage people to pay attention to the ratings, the two pictures in each category were selected to be visually indistinguishable in terms of price, technical specifications, and quality. Participants indicated their final decision by a keyboard press.

Each participant chose twice between each pair of products—once in the *description* format and once in the *experience* format. To control for order effects, we randomized the order of the format and (right versus left) placement of products. Participants were not told that they were making the same decision twice (once in a description and once in an experience format), and the order of the two presentation formats was counterbalanced.[2] To further minimize the influence of prior experience, we asked participants to complete a secondary task[3] that took approximately 20 min between the two formats.

Figure 1 shows screenshots of the description and experience rating formats. As on the majority of e-commerce sites, ratings were displayed as stars. In the description format, the distribution of ratings was represented by a bar plot designed to resemble the summary format used on Amazon.com, in terms of color, style, and information presented (e.g., bars and counts in the description format). Each bar plot consisted of a total of 100 ratings. The total number of ratings of each star value was specified next to the bars. Participants were free to study the bar plot for as long as they wanted before making a final decision. In the experience format, participants sampled consumer ratings sequentially. They pressed a blue or a green key to choose one or the other option and were shown a randomly sampled consumer rating for that product, displayed as a number of stars. There were no constraints in terms of time, number of samples, or sampling sequence. The ratings were randomly drawn with replacement from the underlying hidden distribution of ratings, which was identical to that presented in the description format. Participants indicated when they were ready to stop sampling. Once sampling was terminated, they were asked to make their final choice.

Figure 2 displays an example pair of the distributions employed. In every pair of options, one was clearly unimodal (A). The other option (B) was bimodal and followed the J-shaped pattern described in Hu et al. (2009). These distributions allowed us to study the psychological impact of rare ratings (refer to APPENDIX B for a full table of the choice problems used). For example, based on the complete distribution of ratings in Figure 2, option A has the higher mean rating. However, assuming the rare ratings at the most negative end of option B has little or no impact—because they are not sampled, undersampled, or not recently sampled—then, option B will have the higher experienced mean and may thus be preferred over option A. For all pairs of distributions, it holds that *not* choosing the higher objective mean (HOM) is consistent with underweighting rare product ratings. Put differently, one option always represented the (objectively) higher mean rating; the other option represented the (objectively) higher median rating.

In addition to sampling and choice data, we collected information on the perceived difficulty of a choice. Specifically, participants rated the difficulty of each choice on a scale from 1 (*very easy*) to 5 (*very difficult*).

RESULTS

Two of the 10 distribution pairs were incorrectly specified in our automated protocol for a substantial part of the data collection. The following results are therefore based on only eight of the 10 product choices per format.

[2]The order in which participants worked through the two formats did not affect either choice proportions (description, $t_{61} = 0.93$; $p = .355$; experience, $t_{61} = 1.34$; $p = .185$) or average sample sizes ($t_{61} = 1.52$, $p = .133$).

[3]The secondary task was the automated operation span task developed by Unsworth, Heitz, Schrock, and Engle (2005). We chose this task for two reasons. First, it is a rather long task (~20 min), making carry-over effects from one format to the other relatively unlikely. Second, one previous study has reported a relationship between working memory capacity and sample size (Rakow et al., 2008). We investigated whether this finding could be replicated using a similar working memory task. However, we found no relationships between operation span, as follows: (i) sample size ($r = .04$); (ii) switch rate ($r = -.07$, refer to Hills & Hertwig, 2010); or (iii) subsample size ($r = -.02$).

DOI: 10.1002/bdm

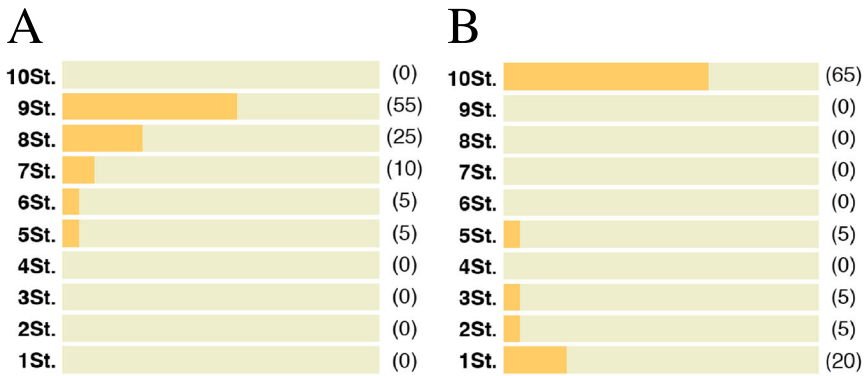

Figure 2. Pair of distributions of consumer ratings. In this example, Distribution A is superior in terms of the mean star rating and therefore more likely to be chosen in the absence of underweighting of rare events. Underweighting rare events, in contrast, should result in favoring Option B (J-shaped distribution).

Is there a description–experience gap in choice based on consumer ratings?

The description and experience rating formats resulted in substantially different choices in relation to the products' objective mean rating. Figure 3 shows that the probability of choosing the product with the HOM rating was about 13 percentage points lower when the choice was based on experience ($M = 65.5\%$, $SD = 19.3\%$) as opposed to description ($M = 78.4\%$, $SD = 24\%$). Thus, even though participants saw the same product options in the experience and description formats, which were both based on the same underlying distributions, the participants chose the HOM option less often when their decisions were based on the experience format, $t(62) = 3.66$, $p < .001$, $d = .59$.[4]

This behavior is consistent with people in experience-based risky choice choosing as if rare events receive less weight than what they deserve according to their objective probability (Hertwig & Erev, 2009; Hertwig et al., 2004). The description–experience gap thus appears to generalize from the domain of monetary gambles to the domain of online consumer choice based on product ratings. Next, we examine to what extent the gap can be explained in terms of small samples and recency.

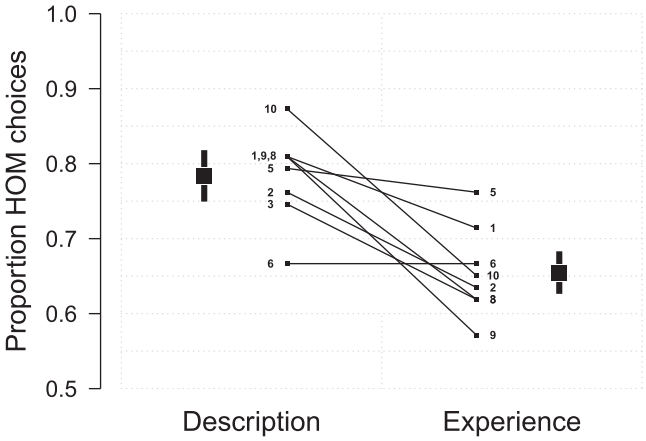

Figure 3. Description-experience gap. Overall proportions of people choosing the higher objective mean separately for description and experience format of consumer ratings are displayed. Lines and numbers represent the decision proportions for the eight problems analyzed. Error bars represent standard error of the mean.

Small samples

Probably, the most important factor in the gap between the description and experience formats is limited search in the experience format (Hertwig & Erev, 2009). Small samples reduce the likelihood of encountering rare ratings (be they positive or negative) and thereby reduce their impact.[5] The

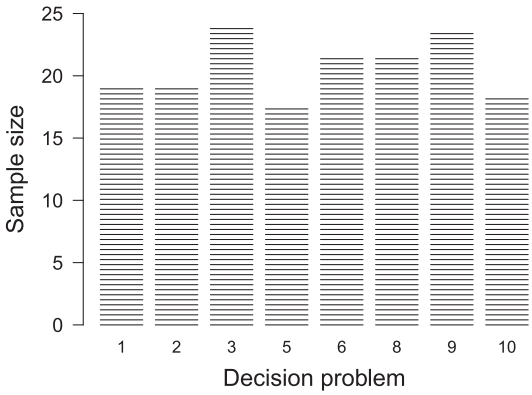

Figure 4. Sample size per decision problem.

average sample size per decision problem varied between 18 and 24 (Figure 4), with a mean of 20.67 ($SD = 2.46$). These numbers are similar to but slightly larger than the sample sizes reported in other computer-based studies of decisions from experience (e.g., Hertwig & Pleskac, 2010; Hertwig et al., 2004). One possible reason for this small increase in sample size is the absence of sure options, which are usually explored less extensively (Lejarraga, Hertwig,

[4]Mixed-effects analyses revealed that the inclusion of a fixed problem factor did not improve the prediction of choices in either the experience or the description format (likelihood-ratio test; experience, $X^2_7 = 7.45$, $p = .38$; description, $X^2_7 = 12.63$, $p = .08$).

[5]The likelihood of experiencing a rating can be understood in terms of the proportion of people that observe the rating less frequently than expected or not at all. Thus defined, a reduced likelihood can be qualified via the shape of the sampling distribution: A right-skewed sampling distribution implies a higher proportion of people experiencing a rating less often than the expected value, and vice versa.The sampling distribution for the number of occurrences of any outcome is governed by a binomial distribution and its skewness is calculated as $\frac{1-2p}{\sqrt{n^* p^* (1-p)}}$. This term is positive (right skewed) for all $p < .5$ and increases with smaller ps and smaller ns. Hence, smaller sample sizes reduce the likelihood of experiencing rare ratings.

DOI: 10.1002/bdm

& Gonzalez, 2012). Critically, the observed sample sizes are sufficiently small to render it possible for small sample size to have a direct impact on choice. For illustration, with 20 draws spread across both products, the chances of experiencing each of five possible star ratings in both options (the maximum per option in this study), assuming an equal distribution of ratings (each rating has a likelihood of 20%), are about one in five. Thus, small sample sizes could easily influence choice.

Indeed, small samples had a direct impact on the final choice. Taking more samples increased the likelihood that participants would choose the option with the HOM rating (the correlation between mean individual sample size and the proportion of HOM mean choices was $r = .47$, $p < .001$). A mediation analysis showed that the reduced sampling of rare ratings was sufficient to explain the fewer choices favoring the HOM ratings (APPENDIX A). Thus, consistent with our predictions, one explanation for the description–experience gap is that participants were content with relatively small samples of ratings in the experience format, thus either missing the rare ratings or experiencing them less frequently than expected. This led participants in experience-based formats to make choices as if they were underweighting rare ratings.

Why do people content themselves with relatively small samples of information that is essentially free? One possible explanation is the *amplification effect* (Hertwig & Pleskac, 2010): Small samples amplify the perceived difference between the expected mean earnings associated with the payoff distributions, thus making the options more distinct and choice easier. The same argument applies to distributions of consumer ratings. Consistent with the amplification effect, we found that our participants' evaluations of choice difficulty were substantially correlated with sample size (subject level: $r = .39$, $p = .002$). Specifically, avid searchers judged decisions to be more difficult than did frugal searchers. However, a within-participant comparison between the experience and description formats revealed that choice difficulty for frugal searchers was not attenuated as has been observed by Hertwig and Pleskac (2010). Following the original Hertwig and Pleskac analysis, we analyzed perceived difficulty as a function of a median split on sample size and the different formats. Frugal searchers judged their experience-based choices to be as easy as those made in the description format, $t(30) = 1.59$, $p = .12$. Avid searchers, in contrast, judged their experience-based choices to be significantly more difficult than those made in the description format, $t(31) = 4.57$, $p < .001$. However, the interaction failed to reach significance, $F(1,120) = 0.4$, $p = .85$. Overall, drawing higher numbers of samples nevertheless appeared to be associated with decreased ease of making decisions.

Of course, this analysis ignores some important information. It glosses over the stratified nature of the data and neglects the mediating role of actual difficulty (i.e., the difference in experienced means between problems). If the amplification effect works as proposed, then higher levels of search should be associated with increased perceived difficulty, and this association should, in turn, be mediated by the actual difficulty. Alternatively, if difficulty is a mere expression of effort, then taking more or fewer samples should remain related to perceived difficulty even after the inclusion of actual difficulty. To address this issue, we performed a mixed-effects analysis[6] predicting the perceived difficulty by sample size and, in the second step, the final Cohen's *d* based on the experienced outcomes of a given problem as an indicator of actual difficulty. To account for dependent measurements, we included two random intercepts for participants and problems.

Consistent with the amplification effect, we found sample size to be highly associated with perceived difficulty ($b = .57$, $p < .001$). Importantly, this association was only moderately reduced by the inclusion of actual difficulty (partial effect: $b = .49$, $p < .001$). Thus, the effort of sampling appears to contribute to the perception of difficulty. However, the pattern of partial mediation is completed by significant associations both between sample size and actual difficulty ($b = -.29$, $p < .001$), with larger sample sizes being related to smaller differences, and between actual difficulty and perceived difficulty ($b = -.16$, $p < .001$; Baron & Kenny, 1986; refer also to APPENDIX A). Thus, in addition to the influence of effort, small samples rendered choices easy.

The influence of recency on consumer choice

In past research, recency has not consistently been observed to affect decisions from experience (refer to Hertwig & Erev, 2009). Did it affect our participants' product choices? We based our analysis on the initial and final samples taken from each option. Sample means were computed for both options' initial and final sampling periods (Figure 5) and compared with respect to their ability to predict the final choices.

Out of 96 cases where the initial and final sampling period suggested different options to be better, participants chose the option that had the higher mean in the most recent sampling period in 72 (74%) cases (sign test, $p < 0.001$). Moreover, a mixed-effects analysis using the means within the *first samples* and *last samples* also revealed a much higher impact for the mean difference in the last samples (*odds ratio* = 17.27, $p < .001$) than that for the mean difference in the first samples (*odds ratio* = 2.31, $p < .001$).[7] Recency thus appears to have played an important role in product choice based on online consumer ratings.

DISCUSSION

Social information in the form of consumer ratings is a driving factor behind online consumer choice. We

[6]Mixed-effects analyses were performed using the R packages *lme4* and *lmerTest*. Degrees of freedom for the Gaussian linear models were estimated using Satterthwaite's approximation, the default method in *lmerTest*. For better comparison, all predictors were standardized.

[7]This analysis was based on 296 of 504 decisions in which the participants switched at least twice between the options. We also tested whether this effect was moderated by differences in the number of samples in the *first samples* (average length = 12.4) and *last samples* (average length = 9). However, the inclusion of two variables representing the number of samples left the effects unchanged and did not result in improved model fit, $X^2(2) = 4.44$, $p = .11$.

DOI: 10.1002/bdm

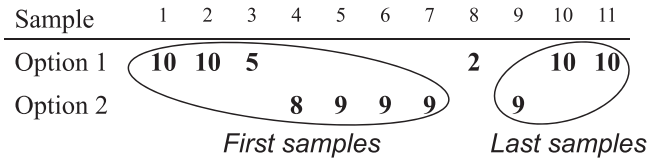

Figure 5. Illustration of *first samples and last samples*. *First samples* include all samples prior to the second switch, and *last samples* include all samples beyond the second to last switch.

investigated the extent to which recent findings in research on decisions from experience in the domain of monetary gambles generalize to choice based on online consumer ratings. Our results suggest that the domain of online consumer choice may be subject to some of the same information-format dependence as observed in risky choice. There is a profound difference between making choices based on a summary "descriptive" format of online consumer ratings and making choices based on sequential sampling from individual consumer ratings, even when the underlying distributions of the ratings are the same (Hertwig & Erev, 2009).

Our results further demonstrate that factors previously proposed to contribute to the description–experience gap may apply more generally. Specifically, we observed three contributors to the description–experience gap in choice based on online consumer ratings: First, people perceived choices to be easier when they took smaller samples (refer to Hertwig & Pleskac, 2010). Second, small sample sizes reduced the likelihood of participants experiencing rare information, leading them to make choices as if they underweighted rare ratings. Third, participants were clearly influenced by the recency of sampled information (Hertwig et al., 2004)—again, leading them to make choices as if they underweighted rare ratings. In sum, the full set of core findings on the description–experience gap persisted in a (hypothetical) online consumer choice scenario in which the outcome distributions were more complex than in previous investigations of the description–experience gap. This not only opens up many new directions for future research but also has specific implications for e-commerce.

In particular, the format dependence of the impact of infrequent ratings is of great importance for e-commerce. As noted by Hu et al. (2009), the majority of consumer rating distributions are J-shaped, with many favorable ratings and few unfavorable ones. Our findings indicate that this will lead people to have lower expectations of consumer goods when looking at summary description-based formats than when perusing individual ratings or entries (but refer to Ert, 2005). Administrators of e-commerce sites can potentially use these findings to foster more informed consumer choice and consumer satisfaction by making sure that consumers always have access to full summary descriptions. Further, the observed recency effect illustrates the relevance of presentation order of consumer ratings. Finally, our findings are relevant for the growing problem of separating truthful from fabricated reviews (Streitfeld, 2013). If fake ratings are both extreme *and* rare, then the use of the experience format would naturally undermine their influence in much the same way as a trimmed mean reduces the influence of strategic scoring in sports competitions (refer to Bamberger, Erev, Kimmel, & Oref-Chen, 2005).

Further, there is a rich set of findings in research on decisions from experience involving risky choice that appears relevant to research on the psychological impact of online product reviews. For instance, it has been demonstrated that the amount of information search substantially varies with factors such as the decision maker's affective state (Frey, Hertwig, & Rieskamp, 2014), the value of the options (Hau, Pleskac, Kiefer, & Hertwig, 2008), the choice domain (i.e., gain versus loss; Lejarraga et al., 2012), and the influence of prior sampling from larger or smaller set sizes (Hills et al., 2013). Another important finding is that the way people search for information in terms of switching between options (or distribution of ratings) foreshadows the decision strategies that people appear to use (Hills & Hertwig, 2010)—providing another potential explanation for the description–experience gap. Specifically, it has been shown that people who often switch between options in the sampling period do not maximize the mean outcome but rather tend to choose an option that is better "most of the time." Finally, in light of the inconsistency of previous findings on recency effects (e.g., Rakow et al., 2008; Ungemach et al., 2009), the pronounced recency effect observed here suggests problem complexity (e.g., number of distinct outcomes/ratings) as a potential moderator of recency in experience-based choice.

Of course, we should emphasize that this first investigation does not reflect the true complexity of e-commerce sites. Most importantly, the majority of sites (e.g., Amazon.com, Tripadvisor.com, etc.) allow consumers to peruse ratings in combination with written reviews. These range from largely uninformative brief statements ("great book") to reviews providing valuable assessments of a product and its properties. Our investigation cannot account for this or for other sources of information (e.g., ratings of the helpfulness of a review, full profiles of reviewers, and total number of ratings). All of these dimensions can and should be addressed in subsequent studies.

Last but not least, we should emphasize that our study—based on hypothetical product reviews and incentivized, but ultimately hypothetical, choices between pairs of consumer products—cannot approximate the rich motivational structure of actual consumer choice. The goals of people buying consumer products of the type used here (e.g., camcorders) will differ from those of our participants. First, a consumer may focus on a single product (rather than two or more

DOI: 10.1002/bdm

products) and is likely to compare products along numerous potentially incommensurable dimensions. Second, as a consumer typically purchases only one, say, camcorder, he or she may aim to minimize the maximum loss (the purchase of a "lemon") or to satisfy an aspiration level for each purchase. In contrast, in our study implementing 10 choices, a bad outcome in one choice can be compensated by a good outcome in another; hence, the participant can aggregate the risk over choices bracketed together (Read, Loewenstein, & Rabin, 1999). Therefore, the robustness of the present results should next be tested in settings with real product ratings, real consumer products, and real choices. Notwithstanding these issues, however, it is worth noting that the description–experience gap obtained in monetary gambles, and replicated here, has also been found in (hypothetical) choices in which people relied on a minimax heuristic (thus avoiding the worst possible outcome), namely, in choices between drugs with different uncertain side effects (Lejarraga, Pachur, Frey, & Hertwig, 2014). Furthermore, individual choice problems in a collection of problems are often played as if they were faced in isolation (Wulff, Hills, & Hertwig, 2014). These results raise the possibility that the gap between choices in the laboratory and consumer choices is perhaps smaller than it might first appear.

The aim of this article has been to relate two hitherto unrelated lines of research on human choice, namely, online consumer choice and risky choice between monetary gambles. The literature on risky choice has produced a large body of experimental findings and theoretical explanations. We found that some key findings on the description–experience gap in risky choice generalize to online consumer choice. This raises the promising and fruitful possibility that other effects observed in research on experience-based and description-based risky choice may also generalize to consumer choice. If so, human choice across different domains may, to some extent, follow the same regularities.

APPENDIX A

The impact of sample size on higher objective mean (HOM) choices was predicted to be mediated by experiencing versus not experiencing rare events. To test this prediction, we performed a mediation analysis on the trial level with the percentage of possible distinct ratings experienced as the mediator. Specifically, we specified a mixed-effects model via the lmer and glmer functions in the R package *lme4*, with random subject intercepts and standardized variables. We found that sample size significantly predicted HOM (*odds ratio* = 1.32, $p = .011$) and the percentage of distinct ratings experienced ($\beta = .53$, $p < .001$), with higher sample sizes leading to more HOM and the observation of more distinct ratings (Figure A1). Thus, two of the necessary conditions in Baron and Kenny's steps for mediation (Baron & Kenny, 1986) are fulfilled. The third condition postulates that the size of the direct effect of the independent variable (sample size) on the dependent variable (HOM) either drops substantially after the inclusion of the mediator (partial mediation) or vanishes completely (full mediation). We found the latter. The effect of sample size on HOM vanished entirely (*odds ratio* = 1.02, $p = .865$) when we controlled for the percentage of distinct ratings experienced. Thus, the effect of sample size on HOM was fully mediated by the percentage of distinct ratings experienced.

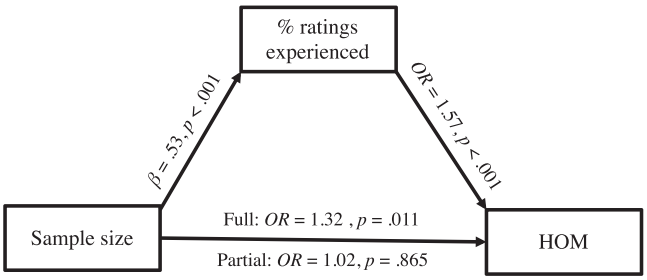

Figure A1. Mediation analysis. "Full" indicates the effect of sample size on higher objective mean (HOM) choices without controlling for percentage of distinct ratings experienced, whereas "Partial" indicates the effect when the mediator is accounted for.

APPENDIX B

Table B1. Choice problems (P) employed in our investigation

	P1		P2		P3		P4		P5		P6		P7		P8		P9		P10	
Stars	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H
1	.2					.65			.35		.05	.05	.25		.1			.6	.3	
2	.05				.9		.25					.65	.05			.1	.6	.05	.05	
3	.05			.7	.05		.05				.7		.05			.6	.35	.05		
4			.9			.05					.15			.05	.8					
5	.05	.05	.1					.1			.1			.05						.05
6		.05						.05		.15				.1			.05			.05
7		.1				.1				.85				.75						
8		.25							.65				.55	.05						.05
9		.55				.15	.05	.8					.1			.05		.15		.85
10	.65			.3	.05	.05	.65	.05				.3				.25		.15	.65	

Note: Entries to left and right of the shaded lines denote the relative frequencies/probabilities of the 1 to 10 star values. Problems 3 and 7 were miscoded for the first 13 of the 63 participants; the numbers displayed correspond to the problems seen by the remaining 50 participants. In order to remedy this mistake, we restricted the analyses to the other eight problems.

DOI: 10.1002/bdm

ACKNOWLEDGEMENTS

We thank Yvonne Bennett and Susannah Goss for editing the manuscript and the Swiss National Science Foundation for a grant to the second author (100014_130397) and the third author (100014–126558).

REFERENCES

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Bamberger, P. A., Erev, I., Kimmel, M., & Oref-Chen, T. (2005). Peer assessment, individual performance, and contribution to group processes: The impact of rater anonymity. *Group & Organizational Management*, *30*, 344–377.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Chen, P., Wu, S., & Yoon, J. (2004). The impact of online recommendations and consumer feedback on sales. *ICIS 2004 Proceedings*, 711–723.

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

Clemons, E. K., Gao, G., & Hitt, L. M. (2006). When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of Management Information Systems*, *23*, 149–171.

Darley, W. K., Blankson, C., & Luethge, D. J. (2010). Toward an integrated framework for online consumer behavior and decision making process: A review. *Psychology & Marketing*, *27*, 94–116

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, *49*, 1407–1424.

Dellarocas, C., Zhang, X., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, *21*, 23–45.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems*, *45*, 1007–1016.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., Hertwig, R., Stewart, T., West, R., & Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*, 15–47.

Ert, E. (2005). Seeing is believing: The positive and negative effects of free sampling. *Unpublished Master's Thesis*, Technion, Israel.

Frey, R., Hertwig, R., & Rieskamp, J. (2014). Fear shapes information acquisition in decisions from experience. *Cognition*, *132*(1), 90–99.

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, *21*, 493–518.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Science*, *13*, 517–523.

Hertwig, R., Hoffrage, U., & Sparr, R. (2012). How estimation can benefit from an imbalanced world. In P. M. Todd, G. Gigerenzer, & the ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world* (pp. 379–409). New York, NY: Oxford University Press.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*, 225–237.

Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience: Do our patterns of sampling foreshadow our decisions? *Psychological Science*, *21*, 1787–1792.

Hills, T. T., Noguchi, T., & Gibbert, M. (2013). Information overload or search-amplified risk? Set size and order effects on decisions from experience. *Psychonomic Bulletin & Review*, *20*, 1023–1031.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *The American Economic Review*, *92*, 1644–1655.

Horrigan, J. B. (2008). *Online shopping*. Washington, DC: Pew Internet Life & American Project Report.

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*, *52*, 144–147.

Johnson, E. J., Bellman, S., & Lohse, G. L. (2003). Cognitive lock-in and the power law of practice. *Journal of Marketing*, *67*, 62–75.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–292.

Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition*, *124*, 334–342.

Lejarraga, T., Pachur, T., Frey, R., & Hertwig, R. (2014). Decisions from experience: From monetary to medical gambles. Manuscript submitted for publication.

Liu, Y. (2006). Word-of-mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, *70*, 74–89.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488.

Peterson, R. A., & Merino, M. C. (2003). Consumer information search behavior and the Internet. *Psychology & Marketing*, *20*, 99–121.

Punj, G. (2012). Consumer decision making on the web: A theoretical analysis and research guidelines. *Psychology & Marketing*, *29*, 791–803.

Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experienced-based choice. *Organizational Behavior and Human Processes*, *106*, 168–179.

Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making*, *14*, 1–14.

Read, D., Loewenstein, G., & Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, *19*, 171–197.

Savage, L. J. (1954/1972). *The foundations of statistics*. New York, NY: Dover.

Streitfeld, D. (2013, September 22). Give yourself 5 stars? Online, it might cost you. *The New York Times*. Retrieved from http://www.nytimes.com/2013/09/23/technology/give-yourself-4-stars-online-it-might-cost-you.html?_r=0

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, *16*(3), 167–171.

U.S. Census Bureau. (2009). Annual retail trade report. Retrieved from http://www.census.gov/retail

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, *20*, 473–479.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498–505.

Vokó, Z., Bots, M. L., Hofman, A., Koudstaal, P. J., Witteman, J. C. M., & Breteler, M. M. B. (1999). J-shaped relations between blood pressure and stroke in treated hypertensives. *Hypertension*, *34*, 1181–1185.

Witteman, J. C. W., Grobbee, D. E., Volakenburg, H. A., van Hemert, A. M., Stijnen, T., Burger, H., & Hofman, A. (1994). J-shaped relation between change in diastolic blood pressure and progression of aortic atherosclerosis. *The Lancet*, *343*, 504–507.

Wulff, D., Hills, T. T., & Hertwig, R. (2014). The impact of short- and long-run frames on search and choice in decisions from experience. *Manuscript submitted for publication*.

DOI: 10.1002/bdm

Authors' biographies:

Dirk U. Wulff is a Predoctoral Fellow at the Center for Adaptive Rationality (ARC) at the Max Planck Institute for Human Development in Berlin, Germany. He received his Masters level degree from the University of Marburg, Germany, focusing on cognitive psychology, clinical psychology and statistics.

Thomas T. Hills is a Professor at the Department of Psychology at the University of Warwick in Coventry, UK. He received his PhD in Biology from the University of Utah. His research focuses on information search, information structure, and their combined influence on learning and memory.

Ralph Hertwig is the Director of the ARC at the Max Planck Institute for Human Development in Berlin, Germany. He held positions at the University of Chicago and Columbia University and served as the chair for Cognitive and Decision Science and the Dean of Research at the University of Basel.

Authors' addresses:

Dirk U. Wulff, Max Planck Institute for Human Development, Berlin, Germany.

Thomas T. Hills, University of Warwick, Coventry, UK.

Ralph Hertwig, Max Planck Institute for Human Development, Berlin, Germany.

DOI: 10.1002/bdm