Swayed by the Numbers: The Consequences of Displaying Product Review Attributes

Jared Watson, jwatson@rhsmith.umd.edu, University of Maryland
Anastasiya Pocheptsova Ghosh, anastasiya@email.arizona.edu, University of Arizona
Michael Trusov, mtrusov@rhsmith.umd.edu, University of Maryland

*As this paper is currently under 2nd round review at the Journal of Marketing, please don’t distribute without consent
Prior research has shown independent effects of average product ratings and review volumes in online purchase incidence, but the relative influence of these review attributes is still debated in the literature. In this research, we demonstrate the conditional influences of these review attributes as a function of the valence of average product ratings and the level of review volumes in the choice set. Specifically, we argue that the diagnosticity of review volumes, relative to average product ratings, increases when average product ratings are low or neutral (versus high) and when the level of review volumes is low (versus high). As a result, when consumers choose between options that are the best on only one of the review attributes (ratings or volume), consumer preference shifts from the higher-rated option with fewer reviews towards the lower-rated option with more reviews. We demonstrate this shift in preference in seven studies, elucidate the underlying process by which this occurs, and conclude with the implications for consumers, retailers and brands.

Words: 165/175

Keywords: online product reviews, average product ratings, review volume, online retail, choice
With the rise of internet shopping, product reviews have gained prominence. Indeed, nearly 60% of consumers now say that the average product rating by other consumers is the most important product attribute in their purchase decisions, while over 80% of consumers trust online reviews as much as personal recommendations (BrightLocal 2016). Because online reviews play such a significant role in consumer behavior, marketing academics have tried to understand the processes by which consumers incorporate the information provided by reviews into their judgments and decisions.

Two attributes of online reviews in particular, average product ratings and review volumes, have received significant attention in recent years (Godes and Mayzlin 2004; Chevalier and Mayzlin 2006; Clemons, Gao, and Hitt 2006; Liu 2006; Dellarocas, Zhang, and Awad 2007; Duan, Gu, and Whinston 2008; Chintagunta, Gopinath, and Venkataraman 2010; Zhu and Zhang 2010; Moe and Trusov 2011; Sun 2012; Ho-Dac, Carson, and Moore 2014). In general, the literature concludes that an increase in either the average product rating or review volume of a product results in an increase in sales. Yet, debate still exists regarding the relative influences of these two attributes. In recent years, two meta-analyses arrived at opposing conclusions. Floyd et al. (2014) finds support for the claim that average product ratings are more influential than review volumes. You, Vadakkepatt, and Joshi (2015), in contrast, argues for the greater importance of review volumes.

Similarly, there is no clear consensus on the preferred presentation of average product ratings versus review volumes by online retailers. For example, Apple.com does not display any review information on its search page. Champssports.com displays the average product rating on its search page, but not the review volume. And Amazon.com displays both the average product rating and review volume on the search page for each product. In a preliminary exploration of
the market, we analyzed the review attribute presentations of over 300 websites and found that while 99 out of the 337 websites chose not to display any review information (29%), of the remaining 228 websites, 54% chose to display review volumes on the search page, while the others chose instead to display review volumes on the product page or reviews page.

Moreover, the conditional effects of average product ratings at different levels of review volumes (and vice versa) on consumer decisions has received little attention to date. A few papers have explored the interactive relationship between various aspects of reviews. Chintagunta, Gopinath, and Venkataraman (2010) investigate the aggregate impact of movie reviews on box office sales (considering only average product ratings but not review volumes of competing choice options) and find no interactive effects of average product ratings and review volumes. Chen, Dhanasobhon, and Smith (2008) investigate the disaggregated impact of individual reviews on book sales. They find that reviews which have received high proportions of “helpful” votes by other consumers are more impactful on sales relative to other reviews, and this effect is stronger for less popular (versus more popular) books. Finally, Khare, Labrecque, and Asare (2011) find that ratings dispersion (i.e., the distribution of individual ratings) differentially impacts sales of negatively- and positively-rated products, but only when review volumes are high. A wide ratings dispersion increases evaluations of negatively-rated products, while a narrow dispersion increases evaluations of positively-rated products. Thus, there is some
indication in previous literature that average product ratings might be evaluated differently under different levels of review volumes (and vice versa), yet the precise nature of when such an interactive effect occurs has not been explored. The goal of the current paper is to address this gap in the literature and specify conditions under which the interactive effect of average product ratings and review volume takes place. In doing so, we contribute to the marketing literature and practice in several important ways.

Theoretically, we contribute to the literature on numerical cognition. Prior work has investigated how expanding versus contracting bound numeric scales of a single attribute influence how consumers evaluate products (Bagchi and Li 2010; Pandelaere, Briers, and Lembregts 2011; Monga and Bagchi 2012; Schley, Lembregts, and Peters 2017). We add to this literature by demonstrating how consumers integrate multiple attributes that use different numeric scales. Specifically, we investigate how consumers integrate average product ratings, which are usually clearly bound on a 1- to 5-point scale, and review volumes, which are usually unbound (could be between zero and infinity). While both average product ratings and review volumes act as signals of quality, we argue that they vary in diagnosticity as a function of the natural differences in the scaling of these attributes.

Consistent with prior work on attribute diagnosticity (Feldman and Lynch 1988; Herr, Kardes, and Kim 1991; Purohit and Srivastava 2001), which suggests that consumers weight attributes in their decisions differently as a function of their perceived diagnosticity as signals of product quality, we demonstrate that in the context of product reviews, average product ratings are more likely to be incorporated in consumer judgments as a signal of product quality than review volumes. Importantly, we further outline conditions under which the perceived diagnosticity of review volumes can increase relative to diagnosticity of average product ratings,
leading to joint influences of average product reviews and review volumes in consumers’ choices. Our investigation into these conditions advances the understanding of how consumers incorporate numerical cues on different scales into their judgments.

Our paper also expands prior work on product reviews in the way we examine the interactive influence of average product ratings and review volumes - by investigating consumer choice between multiple competing product alternatives. Whereas prior research has largely addressed the influences of average product ratings and review volumes on preference of individual choice options (see Chintagunta et al. 2010 for a notable exception), in this research, we examine the influence of these attributes on preferences within a choice set. Increasingly, retailers provide consumers with product options within a choice set rather than individual options (e.g., product search pages, “recommended for you” lists, etc.), thus consumers encounter review information for multiple choice options simultaneously. These choice sets are constructed in a variety of ways. Consumers can construct their own choice set, but often it is done automatically as a function of a retailer’s search feature, and retailers can even curate the choice lists themselves. When consumers select a product on Amazon.com, for example, they are also presented a choice set labeled “customers who viewed this item also viewed” (for various examples, see figure 2), along with an assortment of alternative choice options on the focal product’s page. That is, we investigate the joint influence of average product ratings and review volumes when consumers are choosing among multiple competing choice options that vary on these attributes.

-----

Insert figure 2 about here
Within these choice sets, we specifically examine the conditions by which consumers prefer higher-rated, fewer reviews choice options over lower-rated, more reviews choice options as a function of different levels of average product ratings and review volumes. This allows us to investigate the relative diagnosticity of these two attributes and the conditions by which these relative diagnosticities change. Consider the following scenario. Imagine searching for a new blender online. You see two comparable choice options that have the specifications you are looking for. While one choice option has a higher rating but fewer reviews (e.g., 3.5 out of 5.0 based on 8 reviews), the other choice option has a lower rating but more reviews (e.g., 3.2 out of 5.0 based on 64 reviews). What is the relative diagnosticity of average product ratings and review volumes as a signal of product quality? Would these diagnosticities, and ultimately your decision, change if the review volumes were instead 408 and 464, respectively? And what if the average product ratings were 4.5 and 4.2, respectively?

Managerially, these are important questions to study. We argue that such choice sets, where consumers must choose between product options which have higher average product ratings but fewer reviews relative to options which have lower average product ratings yet more reviews are commonplace. For example, imagine two products released in January and June, respectively. The January product has six additional months to accrue reviews, resulting in a higher review volume, but it has technology that is six months older, potentially resulting in lower quality relative to the June product. Or consider a low-quality brand that ran a brief steep price promotion. Due to the promotion, this low-quality option could have more reviews than its high-quality competitors which did not engage in a major price promotion. Furthermore, as
review volumes increase, the product reach is likely expanding to a more heterogeneous population, resulting in lower ratings due to a lack of fit for all customers.

To demonstrate the prevalence of the described tradeoff scenario we analyzed over 2.5 million products, across 24 product categories, and their corresponding choice sets using data collected from Amazon.com. On average, 79% of the product choice sets featured at least one other product (defined by Amazon as “related and also viewed” product list) that is superior on one of the review attributes, but inferior on the other (see the Appendix W1 for full details of the data set and our analyses). Thus, we argue, that studying these types of consumer choices online is not only interesting from a theoretical standpoint, but also has direct practical relevance as these are the decisions faced by consumers on a regular basis. Hence, from a managerial perspective, we have identified three types of stakeholders – consumers, manufacturers/brands, and online retailers – that could be affected and benefit from this study, and discuss implications specific to each group.

In the following sections, we develop our conceptual framework and the hypotheses to test it. We then test our hypotheses in a series of studies, before concluding with the managerial and theoretical implications of our findings, and directions for future research.

**CONCEPTUAL BACKGROUND**

*The Diagnosticity of Attributes as Signals of Product Quality*

Numerous papers have examined how consumers infer product quality from multiple product attributes when making choices (Slovic 1966; Slovic and Lichtenstein 1971; Rao and Monroe 1988, 1989; Richardson, Dick, and Jain 1994; Kirmani and Rao 2000). Slovic and
Lichtenstein (1971) proposed the concept of attribute diagnosticity in the utilization of attributes in choices. They argue that the perceived diagnosticity of any attribute is a function of the degree to which it separates the available choice options on perceived quality. More diagnostic attributes take precedence over less diagnostic ones as inputs into judgments. The accessibility-diagnosticsity framework further argues that the interpretation of inputs, such as attributes, into choice and judgments is context-dependent rather than fixed (Feldman and Lynch 1988; Lynch, Marmorstein, and Weigold 1988). It suggests that the same attribute can have different perceived diagnosticity depending on context. For example, Lynch, Marmorstein and Weigold (1988) demonstrated that when attributes were easily recalled, they were used as inputs for choice, but when the same attributes were difficult to recall, consumers relied on overall evaluations of the choice rather than individual attributes. Thus, it is not the inherent diagnosticity of an attribute that dictates its usage in decisions, but the perceived diagnosticity of that attribute at the moment of choice.

Importantly, such perceived diagnosticity of one attribute can be a function of the valence of the other attributes. For example, Purohit and Srivastava (2001) demonstrated that manufacturer reputation is considered a highly diagnostic cue, while warranty is not. As a result, whether or not warranty is used in product evaluations depends on the valence of reputation: when reputation is positive, a longer warranty improves quality judgments, while when reputation is negative, the length of warranty does not affect judgments. Thus, the level of one attribute affects the perceived diagnosticity of the other, leading to joint effects of these attributes on consumer preferences.

Related work on attribute evaluability further demonstrates that the levels of the same attribute of the competing choice options can similarly change the perceived diagnosticity of the
attribute and affect preferences (Hsee 1996; Hsee et al. 1999; Hsee 2000; González-Vallejo and Moran 2001; Hsee and Zhang 2010). Hsee (1996) asked participants to evaluate two dictionaries: one dictionary had 10,000 entries and was in perfect condition, while the other had 20,000 entries and had a torn cover. When evaluated independently, the former dictionary was preferred, but when evaluated jointly, preference for the latter dictionary increased. Said another way, when evaluated independently, the condition of the dictionary was more diagnostic, but when evaluated jointly, the number of entries was more diagnostic.

Taken together, this literature provides evidence that the diagnosticity of attributes as signals of product quality and their influence on preference is not fixed and consumers often evaluate attributes depending on the values, and availability, of other attributes in the choice set. Next, we apply these frameworks to consumers’ use of average product ratings and review volumes in decision-making.

Diagnosticity of Average Product Ratings and Review Volumes

Prior work suggests that average product ratings and review volumes are both important signals of product quality, and ultimately, can affect sales (Floyd et al., 2014; You, Vadakkepatt, and Joshi 2015). Complimenting this work, we argue that the relative diagnosticity of each of these two attributes in consumer choices is not fixed and may depend on the value of the other attribute. Specifically, we believe that average product ratings are more easily evaluable by consumers as they directly represent the experienced quality by other consumers. Indeed, recent work demonstrates that consumers believe that average product rating is the strongest indicator of products’ objective quality, more so than other quality cues (De Langhe, Fernbach, and Lichtenstein 2015).
Importantly, average product ratings are generally bound by a scale with two end-points (e.g., 1-5 stars), so consumers can easily compare the average rating to the scale endpoints to infer the level of absolute quality. By contrast, review volumes are presented on an unbound scale: while the minimum number of reviews a product can possess is zero, the potential maximum is infinite. We argue that having review volumes unbound, at least on one end of the scale, can make the absolute number of the reviews more difficult to interpret and lowers the perceived diagnosticity of this attribute in consumers’ judgments.

We build this argument based on numerical cognition literature. In general, numbers and calculations that are easier to process positively improve brand evaluations and product promotions (King and Janiszewski 2011). While not directly tested in prior work, this suggests that, similarly, attribute values presented on bound (versus unbound) scales, which are easier to process, would have greater influence on consumer judgments. In one recent study consistent with this view, Chandon and Ordabayeva (2017) found that consumers made better (i.e., more accurate) judgments, when they were estimating decreasing quantities of food (on a scale bound by two end-points: zero and maximum possible quantity) than increasing quantities of food (on a scale bound by only one end-point: minimum possible quantity and unbound by infinity on the other). Prior work on evaluability mentioned earlier (Hsee 1996), can also be viewed through the lens of bound and unbound scales. The number of words in a dictionary is harder to evaluate, because it does not have a well-defined end-point, while the dictionary’s condition has implicit end-points (perfect and completely destroyed). Thus, the findings that dictionary’s condition was easier to evaluate and did not require additional anchor (in the form of joint evaluation), as compared to number of words in the dictionary, supports our view that attributes expressed on
bound scales are easier to evaluate and would be judged as more diagnostic than attributes expressed on unbound scales.

Building on these literatures, we propose that the diagnosticity of average product ratings as a signal of product quality is higher than that of review volumes. Further, similar to prior work on attribute diagnosticity (Purohit and Srivastava 2001), we argue that whether review volumes are incorporated into decisions would depend on the valence of a more diagnostic cue, average product ratings. When making a buying decision, consumers are motivated to avoid acquisition of an inferior choice option to avoid post-decisional regret (Tsiros and Mittal 2000; Zeelenberg and Pieters 2007). This anticipated regret from making an incorrect (i.e., suboptimal) decision often drives consumers to engage in a more comprehensive assessment of the choice options (Bockenholt et al. 1991). Similarly, research on the negativity bias (Ito et al. 1998; Baumeister et al. 2001; Rozin and Royzman 2001) suggests that consumers attend to, and elaborate more, on information about a judgment in the presence of negative information. Thus, we argue that review volume diagnosticity is likely to increase when average product ratings contain negative information (e.g., low or neutral ratings) relative to when they contain positive information (e.g., high ratings).

-----

Insert figure 3 about here

-----

How would consumers use review volumes in their judgments under these conditions? We believe that the difference in review volumes between two choice options would appear
more diagnostic when the level of the volumes in the choice set is low relative to high. Holding the absolute difference in review volumes between choice options constant, this occurs because the difference between review volumes would appear relatively larger in low (versus high) review volumes choice sets. Prior work demonstrates that people attend more to relative differences than absolute differences in values (Thaler 1980; Tversky and Kahneman 1985). For example, people are willing to exert more effort to save $5 on a $15 purchase than on a $125 purchase. This happens due to a steeper slope for smaller values and a shallower slope for larger values of the utility function, as argued by the Prospect Theory (Kahneman and Tversky 1979; Tversky and Kahneman 1985). Thus, the same absolute difference in numbers (e.g., 10) would loom larger, when the numbers are small (20 versus 30) than large (200 versus 210). Applying this sensitivity to relative differences to the context of review volumes, we would expect that, holding the absolute difference in review volumes between choice options constant, having a choice set with low (versus high) review volumes would increase diagnosticity of the review volume attribute.

The proposed change in perceived diagnosticity of review volumes (relative to average product ratings) has direct implications for consumer preferences. In a choice set, where one option has a higher rating but fewer reviews, while another option has a lower rating but more reviews, an increase in the perceived diagnosticity of review volumes would lead to a weaker preference for the higher-rated option. This occurs because when review volumes are low (relative to high), the diagnosticity of review volumes increases, leading to joint influence of average product ratings and review volumes. As such, when the higher-rated choice option also has fewer reviews, it is perceived as superior on one quality attribute and inferior on another quality attribute, weakening the preference for the higher-rated choice option. However, this
effect of low review volumes on preferences would be attenuated when the average product ratings of both choice options are high, as most consumers will engage in less elaborative decision-making and will be less likely to incorporate review volumes into their decision.

Formally, we propose that given a choice set in which consumers face a tradeoff between average product ratings and review volumes:

**H1:** Preference for higher-rated, fewer reviews choice option will be weaker when average product ratings are low or neutral (versus high), and when the level of review volumes is low (versus high).

**H2:** Preference between choice options is mediated by the difference in perceived diagnosticity of average product ratings and review volumes.

**OVERVIEW OF STUDIES**

Study 1 demonstrates the systematic shift in preference between choice options as a function of review volume levels, providing initial support for H1. Studies 2 and 3 test the generalizability of this effect by examining consequential choice and an expanded choice set. Next, Studies 4 and 5 investigate H1 by demonstrating the moderating influence of high average product ratings on the effect of review volumes on preference, while Study 5 also varies the magnitude of difference between average product ratings to demonstrate the robustness of our effect. Finally, Studies 6 and 7 test H2, in that the difference in perceived diagnosticity of average product ratings and review volumes mediates the effect of review volumes on preference, using self-reported weights of attributes and consumers’ visual attention (captured by eye-tracking equipment) on the attributes.
Studies Paradigm

In every study, participants were asked to imagine that they were considering the purchase of a new product and had narrowed their choice set to two comparable choice options (four choice options in Study 3). Participants then saw both choice options side by side, with information about the brand name, price, average product rating, and review volume for each choice option presented beneath the product images. In each choice set (except for Study 3), choice option A always had a higher average product rating with fewer reviews and choice option B had a lower average product rating with more reviews. Other product attributes were not significantly different.

Specific values of average product ratings and review volumes varied between studies and between the choice sets within the studies to extend the generalizability of our results (see Table 1). Review volumes were chosen by selecting values at the lower and upper limits of the perceived average review volumes based on a pre-test ($N = 182$, see Figure 4) in which we had participants classify various review volumes along a continuum from “Far Fewer than Average (1)” to “Far More than Average (7)”.

Insert tables 1 & 2 about here
After viewing each choice set, participants were asked to indicate their relative preference between choice options on a 7-point scale from “Strongly Prefer Option A (1)” to “Strongly Prefer Option B (7)” (except for Studies 3 and 6). This measure anchored preference for the higher-rated, fewer reviews choice option at “1” and the lower-rated, more reviews choice option at “7”. As such, a higher number on this measure would indicate a weaker preference for the higher-rated, fewer reviews choice option. Figure 5 provides an example of the stimuli participants would view.

The objective of Studies 1-3 was to test H1 by varying different levels of review volumes. We proposed that when consumers face choice sets with low or neutral average product ratings, their preference for the higher-rated, fewer reviews option would be weaker when the choice set had low review relative to high review volumes (keeping the absolute difference in review volumes between options constant between conditions).

**Study 1: Varying Review Volume Levels**

*Participants and design.* 250 participants (M<sub>age</sub> = 31.35; 31% female; Amazon mTurk sample; $0.50 payment) were randomly assigned to one of five review volume levels conditions: low (e.g., 8 vs. 64 reviews), moderate (e.g., 72 vs. 128 reviews), moderately high (e.g., 201 vs.
257 reviews), high (e.g., 456 vs. 512 reviews), or control (i.e., review volumes absent), in a between-subjects design. Within-subject, each participant viewed five product choice sets.

Choice set. For each of the five choice sets (headphones, microwaves, coffeemakers, speaker systems, and lounge chairs), participants would view two products which were nearly identical with the exception of their average product ratings and review volumes (see Table 1). In the control condition, review volumes were not displayed.

Results

A 5 (review volume levels: low, moderate, moderately high, high, control) x 5 (product category: headphones, microwaves, coffeemakers, speaker systems, lounge chairs) repeated-measures ANOVA on preference yielded significant main effects of volume (F(4, 246) = 11.45; p < .001) and product category (F(1, 246) = 18.54; p < .001), while the interaction was not significant (p > .10). Because of this, we collapsed across the product scenarios factor to simplify the reporting of results, though the same directional pattern held for all products. Supporting H1, planned contrasts demonstrated that preference for the higher-rated, fewer reviews choice option was significantly weaker in the low review volumes condition (M_low = 3.96) compared to all other conditions (M Moderate = 2.93; t(246) = -4.74; p < .001; M Moderately high = 2.83; t(246) = -5.12; p < .001; M High = 2.63; t(246) = -6.15; p < .001; and M Control = 2.81; t(246) = -5.27; p < .001). No other contrasts were significant (p > .10). Importantly, not displaying review volumes led to no significant difference in preferences relative to when review volumes were high. This is consistent with our prediction that review volumes are less diagnostic, and therefore, less likely to influence judgments relative to average product ratings when review volumes are high (H2).

In the next study, we replicate the demonstrated shift in preference away from the higher-rated option as a function of review volumes with a consequential choice task.
Study 2: Consequential Choice

Participants and design. We intercepted 105 university students in the student union and offered a $5 payment in exchange for completing several unrelated studies. Participants were entered into a raffle (1 in 50 chance of winning) for the choice option for which they indicated greater preference ($42 value). Participants were randomly assigned to one of two review volume levels (low, high) conditions, in a between-subjects design. At the end of this study participants left their email addresses, in the event that they won the raffle.

Choice set. Participants saw a choice set of two blenders which were nearly identical with the exception of their average product ratings and review volumes (see Table 1).

Results. A one-way ANOVA of review volume levels on relative preference between two choice options was significant (F(1,104) = 19.10; p < .001): preference for the higher-rated choice option was lower when review volumes were low relative to high (M_{low} = 4.68, M_{high} = 3.08), replicating our previous finding for the low and high review volume levels.

In the next study, we expand the choice set from two to four choice options. Prior research has demonstrated that large choice sets increase the use of noncompensatory decision strategies (Payne 1976; Johnson and Meyer 1984), such that consumers are more likely to choose choice options that are superior on one of the most important or easiest-to-differentiate attributes rather than incorporating multiple attributes. To test if expanding the choice set will attenuate our effect, we conducted Study 3.
Study 3: Expanded Choice Set

Participants and design. We recruited 144 undergraduate participants (M<sub>age</sub> = 20.91; 50% female) in exchange for course credit. Participants were randomly assigned to one of three review volume levels (low, high, control) conditions, in a between-subjects design. To capture preference among the four choice options, we used a discrete choice measure rather than the relative preference measure used in Studies 1-2. We then calculated the choice share of the higher-rated, fewer reviews choice option across review volume levels. After this, we also assessed the likelihood of choice deferral by asking participants if they were more likely to purchase one of the available choice options, or defer purchase. The need to make tradeoffs between choice attributes of similar importance increases choice difficulty (Chatterjee and Heath 1996; Dhar and Simonson 2003), which makes choice deferral more likely (Tversky and Shafir 1992; Dhar and Nowlis 1999; Etkin and Ghosh 2017). Thus, we expect the rate of choice deferral to be the highest in the low review volumes condition, compared to the review volumes absent condition, where there are no tradeoffs, and compared to the high review volumes condition, where the diagnosticity of average product ratings is higher than the diagnosticity of review volumes. Similarly, to understand the need for additional information when making a choice, we asked participants “How would you classify the amount of information provided?” from 1 (not enough) to 7 (too much). A more difficult tradeoff between important attributes would require more information to aid the decision process, so we wanted to provide additional evidence for our proposition that a choice set with low review volumes would elicit higher need for information.

Choice set. Participants viewed a choice set of four camping lamps, where options were nearly identical with the exception of their average product ratings and review volumes. While
one choice option had the highest rating with the fewest reviews, another choice option had the lowest rating with the most reviews, and two other choice options in the middle were compromise choice options which were neither the highest, nor lowest on either attribute but were superior on one relative to the other compromise choice option (see Table 1).

-----

Insert figure 7 about here

-----

**Results.**

*Choice of the highest-rated, fewest reviews choice option.* A binary logistic regression in which we dummy coded our review volume levels yielded an omnibus effect of review volume levels ($\chi^2(2) = 10.84; p = .004$). When review volumes were low participants were significantly less likely to choose the highest-rated, fewest reviews choice option ($P_{\text{low}} = 49\%$) relative to when review volumes were high ($P_{\text{high}} = 78\%; \chi^2(1) = 7.09; p = .004$) or absent ($P_{\text{control}} = 76\%; \chi^2(1) = 8.24; p = .008$). There was no significant difference in choice in the high and control conditions ($p > .80$).

*Rate of choice deferral.* A binary logistic regression in which we dummy coded review volume levels yielded an omnibus effect of review volume levels ($\chi^2(2) = 6.73; p = .035$). When review volumes were high ($P_{\text{high}} = 53\%; \chi^2(1) = 4.30; p = .038$) or absent ($P_{\text{control}} = 49\%; \chi^2(1) = 6.02; p = .014$) participants were significantly less likely to defer choice relative to when review volumes were low ($P_{\text{low}} = 73\%$). There was no significant difference between high review volume and control conditions ($p > .70$). A higher rate of choice deferral under low review
volumes is consistent with prior work linking choice difficulty with choice deferral (Tversky and Shafir 1992; Dhar and Nowlis 1999; Etkin and Ghosh 2017). Consistent with our theorizing, when the level of review volumes is low (relative to high or absent), the diagnosticity of review volumes increases (H2), creating a more difficult choice involving the tradeoffs, ultimately increasing choice deferral.

Need for additional information. A one-way ANOVA of review volume levels on need for additional information yielded a marginal effect of review volumes (F(2,144) = 2.98; p = .054). Planned contrasts further demonstrated that participants who encountered high review volumes felt they had significantly more information than those who had encountered low review volumes (M_{high} = 3.37, M_{low} = 2.73; t(144) = 2.43; p = .016). Participants from whom review volumes were withheld were not significantly different from either of the other groups (M_{control} = 3.00; p > .15). While low review volumes increased the need for additional information consistent with our expectations, when review volumes were absent, consumers feel no more need for additional information than when the review volumes are high. While not predicted, this result suggests that withholding review volumes from the list of attributes would not negatively impact consumers’ perceptions of the amount of information they are provided with to make a choice.

Discussion of Studies 1-3

Studies 1-3 demonstrated consistent support for H1: as review volumes decreased, preference for the higher-rated, fewer reviews choice option relative to the lower-rated, more reviews choice option weakened. We replicated this effect across multiple product categories, using different levels of review volumes (Study 1), consequential choice (Study 2), and an expanded choice set (Study 3). Finally, providing initial support for (H2), Studies 1 and 3 also
demonstrated that when review volumes were absent, consumer preferences were similar to that when review volumes in the choice set were high. Further, the rate of choice deferral was higher under low review volume levels consistent with our proposition of an increased diagnosticity of this attribute in consumer decisions.

Studies 1-3 demonstrated that consumers use both average product ratings and review volumes in their choices when the more diagnostic cue, average product ratings, is not positive (i.e., average product ratings were neutral). H1 also suggests that in the presence of a positive high diagnostic cue, consumers are less likely to elaborate on the decision process, relying on fewer attributes to make their decisions, and therefore would be likely to place more weight on average product ratings relative to review volumes. Hence, we expect the effect of review volumes on preference between choice options to be attenuated when product ratings are high. To test this proposition, in Study 4 we manipulate the valence of average product ratings, while keeping the difference between ratings of options in the choice set the same. In Study 5, we relax this last assumption and manipulate both the valence of average product ratings and the magnitude of difference between ratings of options in the choice set.

**Study 4: Average Product Ratings Valence**

*Participants and design.* We recruited 433 undergraduate students (M<sub>age</sub> = 20.28; 46% female) in exchange for course credit. Participants were randomly assigned to a cell in a 2 (review volume levels: low, high) x 3 (rating valence levels: negative, neutral, positive) between-subjects design.

*Choice sets.* Participants saw a choice set of two blenders. Choice options were nearly identical with the exception of their average product ratings and review volumes (see Table 1). Specifically, rating valences were manipulated by changing the first digit of the average product
ratings for both choice options. Thus, the negative condition presented consumers with 2.x choice options, the neutral condition presented 3.x choice options, and the positive condition presented 4.x choice options.

Results and discussion.

A 2 (review volume levels: low, high) x 3 (rating valence levels: negative, neutral, positive) ANOVA on preference yielded main effects of volume (F(1, 427) = 17.26; p < .001) and valence (F(2, 427) = 12.68; p < .001), qualified by the predicted interaction (F(2, 427) = 3.58; p = .029). Planned contrasts provided support for H1: as review volumes decreased, preference for the higher-rated, fewer reviews choice option decreased when average product ratings were negative (M_{low} = 4.15, M_{high} = 3.32; F(1, 427) = 9.77; p = .002) and neutral (M_{low} = 4.38, M_{high} = 3.38; F(1, 427) = 14.28; p < .001), replicating prior studies. Yet, when the average product ratings were positive, the effect of review volumes on preference was attenuated (M_{low} = 3.04, M_{high} = 2.97; F(1, 427) = .07; p > .75). As expected, the influence of review volumes on consumer preferences is attenuated when the more diagnostic cue, average product ratings, is positive.

The next study investigates the role of the magnitude of difference between average product ratings. Building on prior work that argues that the difference in attributes values between choice options can increase the salience of this attribute in decision-making (Monroe
1971; Thaler 1980; Ross and Creyer 1992), we expect that an increase in magnitude of difference between average product ratings of the options in the choice set would increase preference for the higher-rated product. However, consistent with our prior findings, we believe this effect will be more pronounced when diagnosticity of average product ratings is high. Thus, an increase in the difference between average product ratings in the choice set is more salient to consumers when review volumes are high, and the diagnosticity of average product ratings is high, than when review volumes are low and consumers place relatively less weight on the average product ratings.

Study 5: Difference in Average Product Ratings

This study was designed to test the effect of small versus large differences in average product ratings between choice options in the joint influence of review volume and rating valence on preference. In prior studies, we used relatively small differences between average product ratings (e.g., .2-.4), in this study we expand the magnitude of difference to include .6 and .8. Specifically, we created choice sets with differences between choice options of magnitude of .2 (e.g., 3.8 versus 3.6), .4 (e.g., 3.8 versus 3.4), .6 (e.g., 3.8 versus 3.2), and .8 (e.g., 3.8 versus 3.0). We did so across two different valences of product ratings: neutral and high (e.g., 3.x or 4.x). This resulted in a necessarily unbalanced design because there are more differences of .2 within a level than .8, however, since there was no significant difference in effect within a given distance (for example, a difference of 4.2 and 4.4 led to similar preference as a choice set that had 4.4 and 4.6), we collapsed across those cells for the analysis. Thus, we collapsed across all small difference (.2-.4) and large difference conditions (.6-.8) to generalize our findings across relatively small and large difference in average product ratings.
Participants and Design. We recruited 705 participants (M_age = 35.61; 47% female; Amazon mTurk; $0.50 payment). Participants were randomly assigned to a cell in a 2 (review volume levels: low, high) x 2 (ratings valence levels: neutral, positive) x 2 (magnitude of ratings difference: small, large) between-subjects design.

Choice Sets. Participants saw a choice set of two headphones. Choice options were nearly identical with the exception of their average product ratings and review volumes as described earlier (see Table 1), and preference was measured on 7-point scale as in the prior studies.

Results

A 2 (review volume levels: low, high) x 2 (ratings valence levels: neutral, positive) x 2 (ratings difference size: small, large) ANOVA on preference yielded a significant interaction of review volume levels and difference size (F(1, 697) = 4.12; p = .043), an interaction of review volume levels and rating valence levels (F(1, 697) = 3.47; p = .063), and main effects of review volume levels (F(1, 697) = 40.66; p < .001) and rating valence levels (F(1, 697) = 6.70; p = .01). The main effect of review volume levels demonstrates that preference for the higher-rated, fewer reviews option is lower when review volumes are low, consistent with prior studies (M_low = 3.63, M_high = 2.82). The main effect of valence demonstrates that preference for the higher-rated, fewer reviews option is greater when valence is positive relative to neutral (M_neutral = 3.37, M_positive = 3.08), consistent with Study 4.

Consistent with our expectations, the volume by ratings difference interaction indicates that the effect of the magnitude of the average product ratings difference on consumer preferences depends on review volume levels. When review volumes are low, consumers are less sensitive to the magnitude of differences in ratings between options (M_small = 3.62, M_large = 3.68; p > .85). By contrast, when review volumes are high, and, we argue, the relative diagnosticity of
average product ratings is high, consumers are more sensitive to the magnitude of differences in ratings between options, increasing preference for the higher-rated choice option, when the magnitude of difference between options is large versus small ($M_{\text{large}} = 2.45$, $M_{\text{small}} = 2.98$; $p = .01$).

The valence by volume interaction is consistent with the findings of Study 4. While low review volumes decrease preference for the higher-rated, fewer reviews choice option, the magnitude of this effect is larger in the neutral relative to positive valence condition (neutral valence: $M_{\text{low}} = 3.98$, $M_{\text{high}} = 2.77$; $p < .001$; positive valence: $M_{\text{low}} = 3.32$, $M_{\text{high}} = 2.66$; $p = .012$). Note that in this study, positive valence attenuates, rather than completely eliminates, the influence of review volumes on preferences, as seen in Study 4. This could be a function of marginal differences across the ratings differences, but it is important to note that in both studies, consistent with our predictions, positive valence attenuated the influence of low review volumes, albeit to varying strengths.

-----

Insert figures 9 & 10 about here

-----

Discussion. This study replicated the finding of Study 4 by demonstrating that positive valences attenuate the influence of review volumes on consumer decisions. Importantly, it also tested how a magnitude of difference in ratings between choice options changes consumer preference at different levels of review volumes. Our results suggest that ratings differences are more important to consumers when review volumes are high and when the diagnosticity of this attribute is higher relative to the diagnosticity of review volumes. Yet, consumers rely less on the differences in ratings between choice options, when review volumes are low and diagnosticity of
this attribute increases, again highlighting the robustness of the effect of low review volumes on consumer choices.

Having outlined and demonstrated the conditions by which average product ratings and review volumes jointly influence consumer judgments, the objective of the last two studies is to directly test the underlying process and to provide support for H2: the effect of review volume levels on preference is mediated by the difference in perceived diagnosticity of average product ratings and review volumes in consumer choices.

**Study 6: Mediation via the Difference in Perceived Diagnosticity of Average Product Ratings and Review Volumes**

*Participants and design.* We recruited 183 participants (Amazon mTurk; $0.50 payment). Participants were randomly assigned to one of three review volume levels (low, high, control) conditions between-subjects.

*Choice sets.* Participants saw a choice set of two blenders. Choice options were nearly identical with the exception of their average product ratings and review volumes as described earlier (see Table 1). Similar to Study 3, we used discrete choice as our dependent measure and participants were told that they could choose “Option A”, “Option B”, or “Defer purchase and look elsewhere”. Different from other studies, to test H2, we then asked participants to “indicate the importance of each attribute in making your decision” for the five attributes (image, brand, price, average product rating, review volume) on 7-point scales, as a measure of perceived diagnosticity for each attribute.

*Results*

*Choice of the higher-rated, fewer reviews choice option.* Examining only participants who made a choice between the two options (N = 153), a binary logistic regression in which we
dummy coded our review volume levels yielded an omnibus effect of review volume levels ($\chi^2(2) = 15.07; p = .001$). Consistent with our prior studies, when review volumes were high ($P_{\text{high}} = 71\%; \chi^2(1) = 11.40; p = .001$) or absent ($P_{\text{control}} = 71\%; \chi^2(1) = 11.83; p = .001$) participants were significantly more likely to choose the higher-rated, fewer reviews choice option relative to when review volumes were low ($P_{\text{low}} = 36\%$). There was no significant difference in the high and control conditions ($p > .95$).

Rate of choice deferral. A binary logistic regression in which we dummy coded our review volume levels yielded an omnibus effect of review volume levels ($\chi^2(2) = 9.42; p = .009$). The dummy coding demonstrated that when review volumes were high ($P_{\text{high}} = 13\%; \chi^2(1) = 4.64; p = .031$) or absent ($P_{\text{control}} = 8\%; \chi^2(1) = 7.60; p = .006$) participants were significantly less likely to defer choice relative to when review volumes were low ($P_{\text{low}} = 29\%$), replicating results of Study 3. There was no significant difference in choice deferral between the high and control conditions, consistent with earlier studies ($p > .40$).

Perceived diagnosticity of attributes. As expected, there were no significant differences across conditions of the perceived diagnosticity of product image, brand, or price ($p > .10$). Next, we computed a difference score of the perceived diagnosticity of average product ratings and review volumes to demonstrate the changing relative perceived diagnosticities as a function of review volumes. Thus, a lower number on the difference score would indicate that the diagnosticity of review volume is closer to the diagnosticity of average product rating. A one-way (review volume levels: low, high, control) ANOVA on the difference in perceived diagnosticity of average product ratings and review volumes yielded a marginal omnibus effect ($F(2, 150) = 4.59; p = .061$). Consistent with H2, the difference in perceived diagnosticity between the review attributes is lower when review volumes are low ($M_{\text{low}} = .26$), relative to
high (M_{high} = .85; t(150) = 2.28; p = .024) or absent (M_{absent} = .75; t(150) = 1.89; p = .061). There was no significant difference between the absent and high review levels (p > .65). In other words, average product ratings are considered significantly more diagnostic than review volumes in the absent and high review volumes conditions, relative to when review volumes are low and two attributes are closer to each other in diagnosticity.

Preference mediation via differences in perceived diagnosticity of review attributes.

A mediation analysis (Model 4; Preacher, Rucker, and Hayes 2007) was used to demonstrate that the effect of review volumes (low versus high) on consumer preference is driven by the differences in the perceived diagnosticity of average product ratings and review volumes. As expected, the model demonstrated that the effect of review volumes on consumer preference was mediated via the difference in perceived diagnosticity of average product ratings and review volumes (B = -.49; 95% confidence interval [-1.22, -.10]).

Discussion

This study provided support for H2 by demonstrating that the effect of review volumes on choice option preference was driven by the difference in perceived diagnosticity of average product ratings and review volumes: as the perceived diagnosticity of review volumes increases (i.e., when review volumes are low) the preference shifts away from fewer reviews, higher-rated options towards the more reviews, lower-rated option. It further showed, consistent with our propositions, that average product ratings are considered more diagnostic than review volumes, but this difference in diagnosticity is attenuated when review volumes are low. Furthermore, it demonstrated that as the perceived diagnosticity of average product ratings and review volumes become closer, choice deferral increases, consistent with the findings of Study 3 and prior work
demonstrating the link between tradeoff difficulty and choice deferral (Tversky and Shafir 1992; Dhar and Nowlis 1999; Etkin and Ghosh 2017).

In the next study, we demonstrate our proposed mediation process via eye-tracking measurements of attention paid to each of the attributes. We have argued that consumers infer different diagnostic values of average product ratings and review volumes as a function of the level of review volumes. Specifically, in choice sets with neutral and low average product ratings, when consumers see that review volumes are low, it signals to them that average product ratings may not be as diagnostic of product quality as when review volumes are high or when no review volume information is displayed. In terms of consumers’ attention when examining review attributes, we would expect that consumers would be more likely to return to re-examine average product ratings, after viewing low review volumes. This happens, we argue, because consumers need to re-evaluate the average product ratings in light of the new important attribute – review volumes. To test this argument, we use eye-tracking measurements to determine not only the time spent on the attributes, but also the sequence of fixations, to determine whether consumers are more likely to return to examine average product ratings after seeing that review volumes are low versus high.

**Study 7: Mediation using Eye-tracking**

**Method.**

*Participants and design.* We recruited 92 undergraduate participants in exchange for course credit. Participants were randomly assigned to one of two review volume levels conditions (low, high) between-subjects. Participants were randomly selected two at a time from a larger sample of research participants to participate in the eye-tracking study. After engaging in a short eye-tracking calibration task, participants followed a similar paradigm to prior studies.
Choice set. Participants saw a choice set of two microwaves. Choice options were nearly identical with the exception of their average product ratings and review volumes as described earlier (see Table 1). Relative preference between choice options was measured on the same 7-point scale as in studies 1-2.

Additional measures. We defined areas of interest (AOIs) as parts of the screen where corresponding product attributes were displayed and measured the number of eye fixations and gaze times for each attribute. Fixations refer to the frequency participants would look at a given attribute, while gaze times refers to the amount of time spent looking at the specific attributes. As expected, there were no significant differences across conditions for fixations or gaze times of product images, brand names, prices, or highlighted information ($p > .10$), so we do not discuss these further.

Results

Relative preference. A one-way (review volume levels: low, high) ANOVA on preference yielded a significant effect ($F(1, 90) = 10.32; p = .002$). Consistent with prior studies, preference for the higher-rated, fewer reviews choice option was weaker when review volumes were low ($M_{\text{low}} = 4.89$) relative to high ($M_{\text{high}} = 3.68$).

Transition matrices. To provide further support for underlying process, we also derive transition matrices from the eye-tracking data. Doing so allows us to demonstrate the probabilities of participants transitioning their attention from one attribute to the next. As we discussed earlier, we argue that low review volumes are perceived to be more diagnostic, relative to high review volumes, and this causes consumers to re-evaluate average product ratings. To demonstrate this, we assessed the differential probabilities of participants shifting their attention from review volumes to average product ratings as a function of the review volume levels.
Consistent with our theory, participants were significantly more likely to return their attention to the average product ratings after viewing review volumes when the review volumes were low relative to high ($P_{\text{low}} = .24$; $P_{\text{high}} = .13$; $z = 3.25$; $p < .01$). This suggests that participants were nearly twice as likely to return their attention to average product ratings when review volumes were low versus high. Importantly, the transition proportions from review volumes to all other attributes were not significantly different across conditions ($p > .10$).

Insert table 3 about here

*Difference in fixation counts for the review attributes.* Because our variable of interest is the difference in attention paid to average product ratings and review volumes, we calculated the difference in fixations between average product ratings and review volumes. A one-way (review volume levels: low, high) ANOVA on the difference in fixation counts yielded a significant effect ($F(1, 90) = 7.05; p = .009$). As expected, the difference in fixations between average product ratings and review volumes was greater when review volumes were low ($M_{\text{low}} = 4.29$ fixations) relative to high ($M_{\text{high}} = 2.11$ fixations). This is consistent with our view that low review volumes cause consumers to re-evaluate the average product ratings attribute, thus increasing overall attention paid to that attribute.

*Difference in gaze times for the review attributes.* A one-way (review volume levels: low, high) ANOVA on the difference in gaze times yielded a significant effect ($F(1, 90) = 10.59; p = .002$). As expected, the difference in gaze times between average product ratings and review volumes was greater when review volumes were low ($M_{\text{low}} = 12.40$ seconds) relative to high ($M_{\text{high}} = 5.78$ seconds). Consistent with our prior finding, consumers seem to pay more attention
to average product ratings when review volumes are low, and we argue that this occurs because
the low review volumes cause consumers to re-evaluate the inferences from average product
ratings.

*Mediation via the difference in gaze times.* We argue that gaze times are a more precise
measure of attention relative to fixations because they quantify the time spent on an attribute. As
such, we demonstrated that consumers are likely to pay more attention to average product ratings
when review volumes are low, and thus, this difference in gaze times would mediate the
influence of review volumes on consumer preference. Using the mediation analysis (model 4;
Preacher and Hayes 2007), we find support for the argument that the difference in gaze times
between average product ratings and review volumes mediates the effect of review volumes on
consumer preference between choice options ($B = .25; 95\%$ confidence interval $= [.04, .62]$).

*Discussion.* This study provided further evidence for the mediating role of differences in
the perceived diagnosticity of average product ratings and review volumes on consumer
preferences. When review volumes are low, it signals to consumers that the average product
ratings may be relatively closer in diagnosticity to review volumes, leading them to re-evaluate
this attribute before reaching a decision. Yet, when review volumes are high, there is a hierarchy
in diagnosticity between two attributes, and consumers can reach a decision faster without re-
evaluating average product ratings.

**GENERAL DISCUSSION**

Across seven studies we find consistent support for our propositions that average review
ratings are more diagnostic cue of product quality than review volumes and that review volumes
can become more influential in consumer decisions when: a) average product ratings are low or
neutral and b) review volume of the choice set is low. This change in attribute diagnosticity leads
to systematic shift in preference away from the higher-rated, fewer reviews option towards the lower-rated, more reviews option. Furthermore, when the diagnosticity of these two review attributes are closest to each other, consumers experience tradeoff difficulty as evidenced by increased choice deferral.

Robustness Checks and Potential Moderators

We conducted another seven studies to check the robustness of the effects presented in the paper. The studies are reported in full in the Web Appendix and outlined briefly next.

A large body of work on numerosity (Burson, Larrick, Lynch 2009; Pandelaere, Briers, and Lembregts 2011; Lembregts and Pandelaere 2013) makes a different argument to the one we test in the paper. This line of research shows that while keeping relative differences between attribute values of choice options constant, expanding the scale in which the values are presented (e.g., warranty described in months versus years) and thus changing the absolute difference, results in differences between attribute values appearing exaggerated on expanded scale (e.g., warranty described in months). By contrast, in our paper we keep the absolute difference between choice options the same, while changing the relative difference between attribute values by changing the scale. We believe that both effects might exist in the context of product reviews, but it is unclear a priori which one would have a stronger effect on consumer judgments, since prior work on numerosity has not studied its effects in the context of multiple numerical attributes, as we did. Indeed, in an additional study conducted (Appendix W2), we compared the influence of an increase in relative difference between choice options on review volume attribute (as in our studies) to an increase in absolute difference between choice options on review volume attribute (as demonstrated by prior work). This study provided initial evidence that both effect exists, yet at varying degrees of strength.
Next, we tested the role of ratings skew in moderating the influence of review volumes on interpretation of average product ratings. Khare et al. (2011) demonstrated that skew was only influential when review volumes were high when evaluating single options. In this study (Appendix W3), we find both the effect of review volumes and the effect of skew, demonstrating that both effects exists but do not interact with each other when evaluating choice sets. In another study (Appendix W4), we tested whether our effect is attenuated by sequential versus simultaneous presentation of choices. While prior research has demonstrated that presentation mode can influence how consumers process attributes (Schley, Lembregts, and Peters 2017), we show that this has no influence on our effects. This provides greater external validity to our findings as it demonstrates that the joint effect of average product ratings and review volumes can be observed under either presentation mode.

We conducted additional studies that tested the relative effect of other cues of product quality in addition to review volume and product ratings. Specifically, we tested (1) the effect of a popularity cue (e.g., “best-seller” label), (2) the effect of versioned products (e.g., a 2015 versus 2014 model), (3) the effect of an explanation for low review volumes (e.g., a “new arrival” product), and (4) the effect of expert reviews on consumers’ preferences between options (see Appendices W5-W8). These studies replicated our main effect in the presence of these alternative cues of product quality.

Finally, we used ComScore data to provide some external validity to our findings. Without the individual purchase along with the choice sets of each website session, it was impossible to directly replicate our findings in this dataset. Yet, the ComScore data did allow us to provide some support for the strategy of attenuating the role of review volumes on low volume retailers. We assessed how various website metrics (e.g., stickiness, transaction conversion rate,
etc.) differed as a function of whether a retailer chose to display review volumes at all, on the search page (i.e., prominent location), or on the search page (i.e., attenuated location). We urge readers to not overweight these analyses as they are highly susceptible to potential confounds without complete information about purchases and choice sets (for more details, see Appendix W9).

**Managerial Implications**

Next, we provide illustrations of how the results of our study can be incorporated into the business practices. We identify three types of stakeholders – consumers, manufacturers/brands, and online retailers – that could be affected and benefit from this research. We explore several business scenarios where information presentation format could make a difference to these groups.

**Implications for Consumers**

Our results show that consumers are more likely to choose a lower-rated product when it has more reviews. We argue that such behavior may lead to suboptimal decisions as, from a statistical point of view, for certain levels of the average product rating and review volume combinations consumers would be better off selecting a higher-rated, but fewer reviews option. To illustrate this point, in Figure 11 we calculate 95% confidence intervals for sample average product rating under various levels of rating dispersion in the sample (from very consistent ratings across reviewers (i.e., standard deviation of 0.2) to very diverse ratings (i.e., standard deviation of 1.0)).

The graph suggests that under certain conditions (i.e. left side of the graph) consumers should be fairly confident (at least, in a statistical sense as there is no overlap in confidence intervals) that the higher-rated, fewer reviews option is a better choice. Nevertheless, in our experiments (studies 1-3) under the same conditions consumers demonstrate weak
preference for that option. On the other hand, with a high review volume, consumers’ choices are consistent with statistical tests of sample average comparisons (Figure 12). The Prospect Theory (Kahneman and Tversky 1979) may explain the observed behavior as the higher variance around the top-rated options may drive consumers’ assessment of that item being a high-risk option, hence the option with more reviews is seen (incorrectly) as a safer choice. We propose that retailers could help consumers correct for this bias by withholding aggregate review information until the number of reviews reaches a sufficient quantity (a practice currently employed by the Apple iTunes store), or by attenuating the perceived importance of review volumes by withholding them to the review details section of a product (a practice currently employed by Champssports.com). In doing so, consumers will be less likely to choose inferior goods simply because they have more reviews.

-----

Insert figures 11 & 12 about here

-----

_Implications for Brands/Manufacturers_

Brands should be concerned about consumers (a) deferring their choices and (b) making suboptimal choices leading to lower product satisfaction. Many brands sell the exact same product through multiple ecommerce stores. For example, Banana Republic apparel can be found on bananarepublic.com but also Amazon.com. Imagine the case of Banana Republic’s own retail portal. The brand should be concerned with choice deferral, as displaying low review volumes increases the likelihood that consumers may shop elsewhere. Furthermore, if consumers are choosing inferior products (e.g., lower quality) due the influence of review volumes, their satisfaction may decrease, resulting in lower repurchase intentions and weaker brand loyalty. So
it is important that retailers consider their potential review volumes in deciding when, where, and how to implement their review system.

Now consider Banana Republic’s Amazon.com page. Here, Amazon.com gets significantly more traffic than Banana Republic, which leads to more reviews for Banana Republic products, yet also increased competition from similar products. While Banana Republic cannot choose whether or not to display aggregate review information for their products on Amazon.com, they should focus on gaining a critical mass of reviews as quickly as possible. Potential avenues for reaching a critical review mass quicker may include running price discounts and using other promotional activities.

*Implications for the Online Retail Platforms*

Being aware of the peculiarities of consumers’ online decision-making process presented in this paper, online retailers may benefit in several respects. First, retailers can improve the user experience with a more sophisticated product search feature. It is popular for retailers to sort products by average product ratings, which often results in a top of the list filled with low review volume items. According to our findings, this may not be beneficial to the retailer. Perhaps a better sorting algorithm would take review volumes into consideration when displaying product search results, which to our understanding, is something that Amazon.com is currently testing.

Retailers should also consider how to improve their recommendation systems to increase customer welfare in addition to the likelihood of purchase. Product recommendation systems (Bodapati 2008) are becoming increasingly critical as product assortments continue to expand. A better recommendation system could direct consumers’ attention to certain products, while also minimizing any tradeoffs that could lead to choice deferral. Because the quality of recommendations has a direct influence on consumer satisfaction with the shopping experience,
good recommendations should also encourage repeat business. The interaction between average product ratings and review volumes of a choice set will have a direct effect on consumers’ purchase behavior and also their satisfaction with the service. It may also behoove small retailers to consider forming an alliance where they use a common platform which aggregates reviews across all retailers. While we suspect this may be a costly and challenging action, it may help retailers increase their review volumes.

To summarize, in this section we discussed a few ideas of how the findings of our study can affect and improve business practices of various online entities. Our goal here is bring to light the problem of the tradeoff between a set of product attributes (now user-generated) that consumers frequently face when shopping online. As we suggest above, understanding and properly managing this process has direct business implications.

Conclusions and Future Research

This research outlines conditions where the diagnosticity of review volumes as cues of product quality increases relative to diagnosticity of average product ratings, potentially leading to suboptimal decisions for consumers, and an increase in choice deferral for brands and retailers. Theoretically, we argue that an inherent difference in the types of scale in which these attributes are presented (bound and unbound) leads to the observed difference in their diagnosticity, and by demonstrating how consumers integrate attributes on both scales into a single judgment, we contribute to the literature in numerical cognition. Furthermore, we provide some clarity to the debate on the relative influence of average product ratings and review volumes (see: Floyd et al. 2014; You, Vadakkepatt, and Joshi 2015) by demonstrating the conditional influences of volume under various valence conditions.
Future research should explore how review volumes affect interpretation of average product ratings presented on expanded (e.g., 0-100%) or more contracted bound scales (e.g., “Thumbs Up/Down” votes). Furthermore, research on strategies to bind review volumes to a scale (e.g., classifying review volumes based on “ideal” or “sufficient” values but not quantifying the actual number) can attenuate the bias created by review volumes. Future research could also explore how the relative versus absolute difference in values across the attributes would affect the difficulty of tradeoffs between the review attributes.

From a managerial perspective, future research demonstrating how consumers interpret the presence versus absence of different product attributes across multiple retailers would be an interesting avenue, though this is ultimately an empirical question to answer. Furthermore, in applying these findings in a field setting, individual retailers could determine the optimal strategy for their website. For example, given the demographics and preferences of a specific retailer’s clientele, would their customerbase prefer one strategy over another? Lastly, while consumers read very few reviews before making decisions (BrightLocal 2016), exploring the tradeoffs made between review content and aggregate review valence and volume is an important future direction. While the aggregate information is more representative of the products, individual reviews may heighten saliency of specific product details, leading to shift in weights of the various attributes in their decisions. For example, a review highlighting a bad service experience may outweigh the aggregate information indicating that a restaurant is rather popular with high quality food. Thus, understanding when consumers want to read reviews in addition to summary information, and the relationship between these two sources of information is also an important future direction.
REFERENCES

Arndt, Johan (1967), "Role of Product-Related Conversations in the Diffusion of a New Product," 
Journal of Marketing Research, 1 (August), 291-295.

Bagchi, Rajesh, and Xingbo Li (2010), "Illusionary Progress in Loyalty Programs: Magnitudes, 

Baker, Andrew M., Naveen Donthu, and V. Kumar (2016), "Investigating How Word-of-Mouth 
Conversations About Brands Influence Purchase and Retransmission Intentions," Journal of 
Marketing Research, 53 (April), 225-239.

Baumeister, Roy F., Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs (2001), "Bad is 
Stronger than Good," Review of General Psychology, 5 (December), 323.

Bockenholt, Ulf, Dietrich Albert, Michael Aschenbrenner, and Franz Schmalhofer (1991), "The 
Effects of Attractiveness, Dominance, and Attribute Differences on Information Acquisition in 
Multiattribute Binary Choice," Organizational Behavior and Human Decision Processes, 49 
(August), 258-281.

Research, 45 (February), 77-93.

https://www.brightlocal.com/learn/local-consumer-review-survey].

Chatterjee, Subimal, and Timothy B. Heath (1996), "Conflict and Loss Aversion in Multiattribute 
Choice: The Effects of Trade-off Size and Reference Dependence on Decision Difficulty," 
Organizational Behavior and Human Decision Processes, 67 (August), 144-155.


*Organizational Behavior and Human Decision Processes, 86* (November), 216-233.


Pandelaere, Mario, Barbara Briers, and Christophe Lembregts (2011), "How to Make a 29% Increase Look Bigger: The Unit Effect in Option Comparisons," *Journal of Consumer Research, 38* (February), 308-322.


Footnotes

1 This preliminary study explored the review attribute presentations for three of the highest grossing online retail categories (apparel, small electronics, consumer appliances; NRF E-commerce Sales 2010) from over 300 of the largest (based on Alexa.com rankings) retailers, providing a conservative estimate of review display as smaller retailers are less likely to have the functionality for review acquisition. For retailers that chose to display review information in our sample, the average review volume was 390 (SD = 1241), while the median was 15 reviews for their most popular products. For comparison, the average review volume of all products on Amazon.com is 88 reviews (SD = 64), while the median is 2 reviews.

2 Paired sample t-tests were conducted to contrast the product categories. Regardless of the review volume level, preference for the higher-rated, fewer reviews choice option of headphones (Mheadphones = 3.25) was marginally weaker than that for coffeemakers (Mcoffeemakers = 2.96; t(250) = 1.93; p = .055) and speaker systems (Mspeakers = 3.00; t(250) = 1.71; p = .088), and significantly weaker than for the chairs (Mchairs = 2.56; t(250) = 4.93; p < .001). Preference for higher-rated, fewer reviews choice option of the coffeemakers was significantly stronger than for the microwaves (Mmicrowaves = 3.25; t(250) = -2.25; p = .025) and chairs (t(250) = 3.16; p = .002). Preference for higher-rated, fewer reviews choice option of the microwaves was significantly stronger than for the speaker systems (t(250) = 2.00; p = .047) and chairs (t(250) = 5.11; p < .001). Preference for higher-rated, fewer reviews choice option of the speaker systems was significantly stronger than for the chairs (t(250) = 3.65; p < .001). While there were minor categorical differences between products, we urge the reader to not read too far into this as it could merely be a function of the stimuli. The important takeaway is that the same systematic shift in preference occurred across the various product categories.

3 We thank the anonymous reviewer for suggesting this study.

4 Additional results of extensive exploration of various levels of rating dispersion and review volumes are available from the authors upon request.
Table 1 – Design and Measures Summary

<table>
<thead>
<tr>
<th>Study</th>
<th>Product</th>
<th>Option</th>
<th>Average Product Ratings</th>
<th>Review Volumes</th>
<th>Reported Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(2.)</td>
<td>(3.)</td>
<td>(4.)</td>
</tr>
<tr>
<td>1</td>
<td>Over-the-Ear Headphones</td>
<td>A</td>
<td>.5</td>
<td>8</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>.3</td>
<td>64</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>Coffee Makers</td>
<td>A</td>
<td>.4</td>
<td>6</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>.0</td>
<td>58</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>Microwaves</td>
<td>A</td>
<td>.5</td>
<td>9</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>.2</td>
<td>62</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>Speaker Systems</td>
<td>A</td>
<td>.4</td>
<td>12</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>.1</td>
<td>74</td>
<td>148</td>
</tr>
<tr>
<td></td>
<td>Lounge Chairs</td>
<td>A</td>
<td>.7</td>
<td>5</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>.4</td>
<td>98</td>
<td>196</td>
</tr>
<tr>
<td>2</td>
<td>Blenders</td>
<td>A</td>
<td>.4</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>.1</td>
<td>64</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>Camping Lamps</td>
<td>A</td>
<td>.2</td>
<td>61</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>.6</td>
<td>22</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>.8</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>.4</td>
<td>43</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Blenders</td>
<td>A</td>
<td>.4</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>.1</td>
<td>64</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>Earbud Headphones</td>
<td>A</td>
<td>.8, .6, .4, .2</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>.6, .4, .2, 0,</td>
<td>57</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>Blenders</td>
<td>A</td>
<td>.4</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>.1</td>
<td>64</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>Microwaves</td>
<td>A</td>
<td>.5</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>.2</td>
<td>62</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: For Study 5, all pairwise comparisons were used for average product ratings where A was greater than B, resulting in 10 different comparisons.
Table 2 – Means Summary across Studies

<table>
<thead>
<tr>
<th></th>
<th>Average Product Rating</th>
<th>Review Volumes</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Valence</td>
<td>Difference</td>
<td>Low</td>
<td>Moderately Low</td>
<td>Moderately High</td>
<td>High</td>
</tr>
<tr>
<td>Study 1 (n = 250)</td>
<td>Neutral</td>
<td>-</td>
<td>3.96</td>
<td>2.93</td>
<td>2.83</td>
<td>2.63</td>
</tr>
<tr>
<td>Study 2 (n = 105)</td>
<td>Neutral</td>
<td>-</td>
<td>4.68</td>
<td>-</td>
<td>-</td>
<td>3.08</td>
</tr>
<tr>
<td>Study 3 (n = 144)</td>
<td>Neutral</td>
<td>-</td>
<td>49</td>
<td>-</td>
<td>-</td>
<td>78</td>
</tr>
<tr>
<td>Study 4 (n = 433)</td>
<td>Negative</td>
<td>-</td>
<td>4.15</td>
<td>-</td>
<td>-</td>
<td>3.32</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>-</td>
<td>4.38</td>
<td>-</td>
<td>-</td>
<td>3.38</td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>-</td>
<td>3.04</td>
<td>-</td>
<td>-</td>
<td>2.97</td>
</tr>
<tr>
<td>Study 5 (n = 705)</td>
<td>Neutral</td>
<td>-</td>
<td>3.98</td>
<td>-</td>
<td>-</td>
<td>2.77</td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>-</td>
<td>3.32</td>
<td>-</td>
<td>-</td>
<td>2.66</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>Small</td>
<td>3.62</td>
<td>-</td>
<td>-</td>
<td>2.98</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>Large</td>
<td>3.68</td>
<td>-</td>
<td>-</td>
<td>2.45</td>
</tr>
<tr>
<td>Study 6 (n = 143)</td>
<td>Neutral</td>
<td>-</td>
<td>36</td>
<td>-</td>
<td>-</td>
<td>61</td>
</tr>
<tr>
<td>Study 7 (n = 92)</td>
<td>Neutral</td>
<td>-</td>
<td>4.09</td>
<td>-</td>
<td>-</td>
<td>3.68</td>
</tr>
</tbody>
</table>

Notes: For studies 1, 2, 4, 5, and 7, “1” indicates a strong preference for the higher-rated, fewer reviews choice option, while “7” indicates a strong preference for the lower-rated, more reviews choice option. For studies 3, and 6, the number indicates the percentage of participants selecting the higher-rated, fewer reviews choice option.
Table 3 – Study 7: Eye-tracking Attribute Transition Matrices by Review Volume Level

<table>
<thead>
<tr>
<th>LOW REVIEW VOLUMES</th>
<th>To Image</th>
<th>To Brand &amp; Price</th>
<th>To Rating</th>
<th>To Volume</th>
<th>To Add'l Info</th>
<th>To End</th>
</tr>
</thead>
<tbody>
<tr>
<td>From Image</td>
<td>68.23%</td>
<td>14.47%</td>
<td>4.51%</td>
<td>2.26%</td>
<td>6.02%</td>
<td>4.51%</td>
</tr>
<tr>
<td>From Brand &amp; Price</td>
<td>8.94%</td>
<td>72.81%</td>
<td>12.23%</td>
<td>2.37%</td>
<td>2.92%</td>
<td>0.73%</td>
</tr>
<tr>
<td>From Rating</td>
<td>3.50%</td>
<td>10.07%</td>
<td>62.58%</td>
<td>19.69%</td>
<td>3.72%</td>
<td>0.44%</td>
</tr>
<tr>
<td>From Volume</td>
<td>2.65%</td>
<td>3.41%</td>
<td>24.24%</td>
<td>49.62%</td>
<td>18.94%</td>
<td>1.14%</td>
</tr>
<tr>
<td>From Add'l Info</td>
<td>9.66%</td>
<td>1.87%</td>
<td>1.71%</td>
<td>2.80%</td>
<td>82.09%</td>
<td>1.87%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HIGH REVIEW VOLUMES</th>
<th>To Image</th>
<th>To Brand &amp; Price</th>
<th>To Rating</th>
<th>To Volume</th>
<th>To Add'l Info</th>
<th>To End</th>
</tr>
</thead>
<tbody>
<tr>
<td>From Image</td>
<td>71.74%</td>
<td>12.89%</td>
<td>3.64%</td>
<td>2.15%</td>
<td>6.12%</td>
<td>3.47%</td>
</tr>
<tr>
<td>From Brand &amp; Price</td>
<td>9.35%</td>
<td>73.28%</td>
<td>12.02%</td>
<td>1.53%</td>
<td>2.48%</td>
<td>1.34%</td>
</tr>
<tr>
<td>From Rating</td>
<td>4.26%</td>
<td>9.31%</td>
<td>64.10%</td>
<td>14.63%</td>
<td>6.12%</td>
<td>1.60%</td>
</tr>
<tr>
<td>From Volume</td>
<td>4.69%</td>
<td>2.17%</td>
<td>13.36%</td>
<td>60.65%</td>
<td>17.33%</td>
<td>1.81%</td>
</tr>
<tr>
<td>From Add'l Info</td>
<td>8.20%</td>
<td>1.64%</td>
<td>1.37%</td>
<td>4.37%</td>
<td>83.47%</td>
<td>0.96%</td>
</tr>
</tbody>
</table>
Figures

Figure 1 – Sample of Online Retail Review Landscape

Proportion of Online Retailers which Display Reviews and Whether Review Volumes are Displayed on Search

- Review Information Absent: 0.29
- Review Information Present: 0.71
- Review Volumes Displayed on Search Page: 0.46
- Review Volumes Withheld from Search Page: 0.54
Figure 2 – Choice Set Examples

Hamilton Beach Power Elite Multi-Function Blender with Glass Jar and Chopper (58149) by Hamilton Beach

$34.00
Product Description
...queries please review http://www.hamiltonbeach.com/products/blenders.html ...

Ninja Professional Blender (BL610) by Shin Nutrition

$76.34
Product Description
...The Ninja Professional Blender 1000 features a sleek design and ...

Customers who viewed this item also viewed
Figure 3 – Conceptual Model

- Review Volumes (RV)
- Average Product Ratings (APR)
- RV Level
- APR Valence
- Difference in Perceived Diagnosticity of APR and RV
- Option Preference
Figure 4 – Perceptions of Review Volumes Relative to “Average” Based on Pre-test
Figure 5 – Example of the Standard Study Stimuli

<table>
<thead>
<tr>
<th>Name</th>
<th>Price</th>
<th>Rating</th>
<th># of Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuisinart® Rotisserie Convection Toaster Oven</td>
<td>$169.98</td>
<td>3.8 / 5.0</td>
<td>408</td>
</tr>
<tr>
<td>KitchenAid® Convection Toaster Oven</td>
<td>$168.99</td>
<td>3.0 / 5.0</td>
<td>454</td>
</tr>
</tbody>
</table>
Figure 6 – Study 1 Means

Relative Preference across Five Review Volume Levels

Prefer for the lower-rated, more reviews choice option

Review Volume Levels

- Low
- Moderate1
- Moderate2
- High
- Control (Absent)

Mean values:

- Low: 3.96
- Moderate1: 2.93
- Moderate2: 2.83
- High: 2.63
- Control (Absent): 2.81
Figure 7 – Study 3 Stimuli

A: Coleman & CFX 3.2 / 5.0 (61 reviews) - $49.99
B: Coleman Sidekick 3.6 / 5.0 (22 reviews) - $50.00
C: Eveready 36 Dynamo 3.8 / 5.0 (4 reviews) - $49.95
D: Stansport 3FMOS 3.4 / 5.0 (43 reviews) - $49.97
Figure 8 – Study 4 Preference Results

The Moderating Role of Valence

Preference for the lower-rated, more reviews choice option

Negative Ratings Neutral Ratings Positive Ratings

Low Review Volumes High Review Volumes
Figure 9 – Study 5 Preference Results: Valence by Volume Interaction

The Moderating Role of Valence

Preference for the lower-rated, more reviews choice option

Neutral Ratings

Positive Ratings

Low Review Volumes

High Review Volumes
Figure 10 – Study 5 Preference Results: Ratings Difference by Volume Interaction

The Moderating Role of Ratings Difference

<table>
<thead>
<tr>
<th>Preference for the lower-rated, more reviews choice option</th>
<th>Low Review Volumes</th>
<th>High Review Volumes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.62</td>
<td>2.98</td>
</tr>
<tr>
<td></td>
<td>3.66</td>
<td>2.44</td>
</tr>
</tbody>
</table>

Small Difference (.2-.4)  Large Difference (.6-.8)
Figure 11 – Confidence Interval Chart for Low Review Volumes

Confidence Intervals for "True" Ratings of Options with Low Review Volumes assuming Various Standard Deviations across the Individual Ratings

- 3.5 / 5.0, 8 Reviews
- 3.2 / 5.0, 64 Reviews
Figure 12 – Confidence Interval Chart for High Review Volumes

Confidence Intervals "True" Ratings for Options with High Review Volumes assuming Various Standard Deviations across the Individual Ratings

- 3.5 / 5.0, 408 Reviews
- 3.2 / 5.0, 464 Reviews
Web Appendices

Appendix W1

Amazon Choice Set Data

To understand just how often consumers are faced with a possible tradeoff between average product ratings and review volumes in choice sets, we analyzed a publicly-available data set from McAuley et al. (2015) which included over 142 million reviews and 2.5 million products, and respective consideration sets, for Amazon products across 24 different categories from May 1996 – July 2014. The data was split into two files, one which featured all reviews for the products, and one which featured the metadata, including the “related and also viewed” choice set options. By parsing these lists together, we were able to reconstruct the consideration sets consumers encountered. We then computed two measures from this data. First, we coded for whether any alternative choice option created a tradeoff scenario (i.e., a flip) with the focal product in which one option had a higher average product rating and fewer reviews relative to the other. We also coded for the frequency of this occurrence in each choice set. Thus, we have a measure of whether a tradeoff occurred in each set (i.e., share of choice sets flipped at least once), and how many options forced a tradeoff in that choice set (i.e., flipped share of the choice sets).

Our analysis revealed that 2,050,549 of the 2,503,422 (79%) choice sets demonstrated at least one tradeoff in the choice set between the focal product and its alternatives. Furthermore, on average, nearly half of the choice set, 8 out of 18 options (47%), had a lower rating but more reviews than the focal product.
Thus, the tradeoffs consumers face seem to be rather common in the online market. In fact, given that this information was based on Amazon data, which generally has more reviews than other online retailers, these findings are a conservative estimate of the state of the market, given the increased variance around product ratings when fewer reviews are available. Thus, for most retailers, the frequency of tradeoffs occurring is likely much greater.

Table W1 – Amazon Tradeoff Data

<table>
<thead>
<tr>
<th>Category ID</th>
<th>Category</th>
<th>Number of Products</th>
<th>Products Flipped* at Least Once</th>
<th>Share of Products flipped at least once (%)</th>
<th>Smallest CS size</th>
<th>Largest CS size</th>
<th>Smallest number of items &quot;flipped&quot; in CS</th>
<th>Largest number of items &quot;flipped&quot; in CS</th>
<th>Average number of items &quot;flipped&quot; in CS</th>
<th>Average &quot;flipped&quot; share***</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Amazon Instant Video</td>
<td>110</td>
<td>83</td>
<td>75%</td>
<td>0</td>
<td>57</td>
<td>12</td>
<td>0</td>
<td>52</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>Apps for Android</td>
<td>6,707</td>
<td>4,258</td>
<td>63%</td>
<td>0</td>
<td>60</td>
<td>17</td>
<td>0</td>
<td>58</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>Automotive</td>
<td>213,414</td>
<td>174,078</td>
<td>82%</td>
<td>0</td>
<td>60</td>
<td>19</td>
<td>0</td>
<td>59</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>Baby</td>
<td>44,517</td>
<td>40,179</td>
<td>90%</td>
<td>0</td>
<td>60</td>
<td>29</td>
<td>0</td>
<td>59</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>Beauty</td>
<td>175,633</td>
<td>150,671</td>
<td>86%</td>
<td>0</td>
<td>60</td>
<td>20</td>
<td>0</td>
<td>58</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>Books</td>
<td>261,207</td>
<td>180,081</td>
<td>69%</td>
<td>0</td>
<td>60</td>
<td>6</td>
<td>0</td>
<td>57</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>CDs and Vinyl</td>
<td>87,410</td>
<td>60,081</td>
<td>69%</td>
<td>0</td>
<td>59</td>
<td>6</td>
<td>0</td>
<td>54</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>Cell Phones and Accessories</td>
<td>72,863</td>
<td>57,815</td>
<td>79%</td>
<td>0</td>
<td>60</td>
<td>16</td>
<td>0</td>
<td>56</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>Digital Music</td>
<td>5,832</td>
<td>3,325</td>
<td>57%</td>
<td>0</td>
<td>56</td>
<td>5</td>
<td>0</td>
<td>55</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>Electronics</td>
<td>177,264</td>
<td>145,417</td>
<td>82%</td>
<td>0</td>
<td>60</td>
<td>19</td>
<td>0</td>
<td>58</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>Grocery and Gourmet Food</td>
<td>109,956</td>
<td>93,351</td>
<td>85%</td>
<td>0</td>
<td>60</td>
<td>19</td>
<td>0</td>
<td>55</td>
<td>10</td>
</tr>
<tr>
<td>13</td>
<td>Health and Personal Care</td>
<td>180,015</td>
<td>156,215</td>
<td>87%</td>
<td>0</td>
<td>60</td>
<td>24</td>
<td>0</td>
<td>57</td>
<td>12</td>
</tr>
<tr>
<td>14</td>
<td>Home and Kitchen</td>
<td>251,964</td>
<td>218,096</td>
<td>87%</td>
<td>0</td>
<td>60</td>
<td>25</td>
<td>0</td>
<td>59</td>
<td>12</td>
</tr>
<tr>
<td>15</td>
<td>Kindle Store</td>
<td>43,037</td>
<td>29,954</td>
<td>70%</td>
<td>0</td>
<td>60</td>
<td>11</td>
<td>0</td>
<td>52</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>Movies and TV</td>
<td>34,569</td>
<td>24,131</td>
<td>70%</td>
<td>0</td>
<td>60</td>
<td>7</td>
<td>0</td>
<td>52</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>Musical Instruments</td>
<td>32,306</td>
<td>25,762</td>
<td>80%</td>
<td>0</td>
<td>60</td>
<td>18</td>
<td>0</td>
<td>55</td>
<td>8</td>
</tr>
<tr>
<td>18</td>
<td>Office Products</td>
<td>74,213</td>
<td>63,361</td>
<td>85%</td>
<td>0</td>
<td>60</td>
<td>24</td>
<td>0</td>
<td>59</td>
<td>12</td>
</tr>
<tr>
<td>20</td>
<td>Pet Supplies</td>
<td>69,424</td>
<td>60,128</td>
<td>87%</td>
<td>0</td>
<td>60</td>
<td>24</td>
<td>0</td>
<td>57</td>
<td>12</td>
</tr>
<tr>
<td>21</td>
<td>Sports and Outdoors</td>
<td>285,373</td>
<td>241,576</td>
<td>85%</td>
<td>0</td>
<td>60</td>
<td>19</td>
<td>0</td>
<td>58</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Tools and Home Improvement</td>
<td>129,065</td>
<td>104,454</td>
<td>81%</td>
<td>0</td>
<td>60</td>
<td>19</td>
<td>0</td>
<td>56</td>
<td>9</td>
</tr>
<tr>
<td>----</td>
<td>----------------------------</td>
<td>---------</td>
<td>---------</td>
<td>-----</td>
<td>---</td>
<td>----</td>
<td>----</td>
<td>---</td>
<td>----</td>
<td>---</td>
</tr>
<tr>
<td>22</td>
<td>Toys and Games</td>
<td>234,300</td>
<td>206,437</td>
<td>88%</td>
<td>0</td>
<td>60</td>
<td>31</td>
<td>0</td>
<td>60</td>
<td>14</td>
</tr>
<tr>
<td>23</td>
<td>Video Games</td>
<td>14,243</td>
<td>11,096</td>
<td>78%</td>
<td>0</td>
<td>60</td>
<td>15</td>
<td>0</td>
<td>59</td>
<td>7</td>
</tr>
</tbody>
</table>

* Two products are being "flipped" if one product has higher average but lower number of ratings compared to another product.

** In this context the term "Consideration Set" denotes the products identified by Amazon.com as "related and also viewed" with respect to the focal product.

*** Share of the products in CS flipped with respect to the focal product.
APPENDIX W2

Relative versus Absolute Differences in Review Volumes

The purpose of this study was to compare the influence of an increase in the relative difference between choice options on the review volume attribute (the goal of the main paper) to an increase in the absolute difference between choice options on the review volume attribute as demonstrated by prior work (numerosity effect; Pandelaere, Briers, and Lembregts 2011; Bagchi and Li 2010). To this end, we created the low review volumes and high review volumes conditions, where the absolute difference is held constant while relative difference changes, similar to previous studies, while also adding a high review volumes condition in which the relative difference is held constant with the low review volumes conditions, while the absolute difference changes. This last condition, therefore, provides a test of numerosity effect observed in prior work but in the context of multiple numerical attributes.

Method

Participants and design. We recruited 153 participants (M_{age} = 36.07; 55% female) from Amazon mTurk in exchange for a $0.50 payment. Participants were randomly assigned to a condition in a 3 (review volume levels: low, high relative, high absolute) between-subjects x 3 (product replicates: BBQ grills, patio furniture, patio umbrella) within-subject mixed design.

Procedure. We used the same procedure as in Study 1. For each product replicate, participants would indicate their relative preference between two choice options that required a tradeoff between a higher-rated, fewer reviews option and a lower-rated, more reviews option. In addition to the low (e.g., 12 vs. 45) and high (e.g., 212 vs. 245) review volume conditions used in past studies (manipulating relative difference), we also included a high review volume (e.g., 212 vs. 795) condition (manipulating absolute difference). Following this, we also measured the need for additional information.
**Results.** A repeated measures ANOVA of review volume levels and product replicates on preference yielded main effects of both product replicates (F(1, 150) = 17.74; \(p < .001\)) and review volume levels (F(1, 150) = 8.40; \(p < .001\)), qualified by a significant quadratic interaction (F(2, 150) = 4.06; \(p = .019\)). To explain this interaction, we first compare the low and high relative difference review volumes conditions, which are most similar to our prior studies.

Consistent with prior results, the repeated measures ANOVA of product replicates and review volume levels yielded significant main effects of product replicates (F(1, 98) = 25.09; \(p < .001\)), and review volume levels (F(1, 98) = 11.94; partial \(\eta^2 = .11\); \(p = .001\)), while the interaction was not significant (\(p > .65\)). Consistent with our previous studies, preference for the higher-rated options was greater when review volumes were high (M\text{relative} = 3.22) relative to low (M\text{low} = 4.03). The main effect of product replicates merely demonstrates that consumer preference for the higher-rated, fewer reviews option is weaker for patio furniture (M\text{furniture} = 4.43) relative to the BBQ grills (M\text{grills} = 3.11; \(p < .001\)) and the patio umbrellas (M\text{umbrellas} = 3.33; \(p < .001\)). There was no significant difference in preference between BBQ grills and patio umbrellas (\(p > .30\)).

By contrast, when examining low and high absolute difference review volumes levels conditions, a repeated measures ANOVA yielded a significant main effect of product replicates (F(1, 102) = 8.63; \(p = .004\)), qualified by an interaction with product replicates and review volume levels (F(1, 102) = 6.85; partial \(\eta^2 = .06\); \(p = .01\)). Importantly, the main effect of review volumes was not significant (\(p > .50\)). Planned contrasts demonstrate that while BBQ grills (M\text{low} = 3.31; M\text{absolute} = 4.02; F(1, 102) = 4.04; partial \(\eta^2 = .04\); \(p = .047\)) exhibit a pattern consistent with a numerosity effect, patio furniture exhibits a reverse pattern (M\text{low} = 4.90; M\text{absolute} = 4.23; F(1, 102) = 2.87; partial \(\eta^2 = .03\); \(p = .093\)), and preference for patio umbrellas
was directionally consistent with a numerosity effect ($M_{\text{low}} = 3.86$; $M_{\text{absolute}} = 4.28$; $F(1, 102) = 4.04$; partial $\eta^2 = .01$; $p > .25$).

**Discussion.** The results of this study suggest that both relative and absolute differences in review volumes can affect consumer preferences in the context of integration of multiple numerical attributes, and gives initial evidence that our proposed effect (effect of relative differences) is stronger under conditions where consumers have to tradeoff two numeric attributes.

**Sample Stimuli**
APPENDIX W3

Ratings Skew

The primary purpose of this study was to test the role of ratings’ skew in moderating the joint effect of review volumes and average product ratings. The skew of ratings provides additional information about the expected quality of a product by informing consumers how different proportions of the consumers felt about a product. Khare et al. (2011) demonstrated that ratings’ skew affected consumers’ judgments only when review volumes were high. However, since Khare et al. (2011) investigated single option choices, effect of skew of product ratings was not tested in a setting where consumers have to tradeoff between review volumes and product ratings, the context of this paper.

Method

Participants and design. We recruited 167 undergraduate students in exchange for course credit. Participants were randomly assigned to a condition in a 2 (review volume levels: low, high) x 3 (option A ratings’ skew: absent, negative, positive) between-subjects factorial.

Procedure. We manipulated review volumes across two product options in the same manner as in prior studies. In addition, the skew of average product ratings was manipulated by reporting the percentage of reviews for each possible rating for Option A (higher-rated, fewer reviews option). In the positive condition, the percentage of reviews was 0%, 22%, 33%, 44%, 0% for 1-5 stars respectively, demonstrating that they had twice as many 4-star reviews as 2-star reviews. In the negative condition, the pattern was reversed (0%, 44%, 33%, 22%, 0%). For all conditions where skew was present, Option B always had an even dispersion of 0%, 33%, 33%, 33%, 0% for 1-5 stars respectively. Next, participants indicated relative preference between two choice options.
Results. A 2 (review volumes) by 3 (skew) ANOVA on relative preference yielded significant main effects of volume (F(1, 161) = 16.64; p < .001) and skew (F(2, 161) = 26.82; p < .001). The interaction was not significant (p > .30). Consistent with prior studies, preference for the higher-rated, fewer reviews option was greater in the presence of high review volumes (M_{high} = 3.30) relative to when the consideration set featured low review volumes (M_{low} = 4.25). Planned contrasts demonstrated that preference for the higher-rated, fewer reviews option was weaker in the presence of a negative skew (M_{negative} = 4.91) relative to when a skew was absent (M_{absent} = 3.62; t(164) = -4.30; p < .001). Furthermore, a positive skew increased preference for the higher-rated, fewer reviews option relative to when a skew was absent (M_{positive} = 2.86; t(164) = -2.53; p = .012).

Discussion. This study demonstrated that ratings skew plays a significant role in consumer preferences, but does not interact with the other review attributes. We argue that this occurs because skew is less diagnostic (i.e., more difficult to interpret) relative to average product ratings and review volumes, and thus, consumers are less likely to use it in their judgments relative to the other attributes. Thus, in the presence of other diagnostic attributes, ratings skew influences consumers but not as a function of review volume levels.

Sample Stimuli
Single versus Joint Evaluation of Choice Options

The purpose of this study was to demonstrate the robustness of the interactive effect of review volumes and product ratings on consumer preference under different presentation modes: sequential (i.e., single) versus simultaneous (i.e., joint). Prior research has demonstrated that presentation mode can influence how consumers process attributes, specifically, attenuating numerosity effect (Schley, Lembregts, and Peters 2017).

Method

Participants and design. We recruited 165 participants (M_age = 41.09; 55% female) from Amazon mTurk in exchange for a $0.50 payment. Participants were randomly assigned to a condition in a 2 (review volume levels: low, high) by 2 (presentation mode: sequential, simultaneous) between-subjects by 2 (product replicates: digital camera, tower fan) within-subject mixed design.

Procedure. We manipulated review volumes across two product options in the same manner as prior studies. In addition, presentation mode was manipulated by presenting the options either simultaneously (as in prior studies) or one at a time. After viewing both options, participants indicated preference between options and relative importance of the attributes, measured as in Study 6.

Results

Option preference. A repeated measures 2 (review volumes) by 2 (presentation mode) by 2 (product replicates) ANOVA on relative preference yielded significant main effects of volume (F(1, 161) = 30.78; p < .001) and product replicates (F(1, 161) = 14.93; p < .001). No other effects were significant (p > .10). Preference for the higher-rated, fewer reviews option was
greater for the tower fan (M_{fan} = 3.46) relative to the digital camera (M_{camera} = 4.16). Consistent with prior studies, preference for the higher-rated, fewer reviews option was greater in the presence of high review volumes (M_{high} = 3.24) relative to when the choice set featured low review volumes (M_{low} = 4.40).

Mediation via attribute diagnosticity. As expected, no significant differences existed across image, brand, and price as a function of presentation mode or review volumes (p > .25). As before, we calculated the difference in the perceived diagnosticity of average product ratings and review volumes to create a difference measure, averaging across product replicates. A one-way ANOVA on the difference measure yielded a significant effect of review volumes (F(1, 161) = 8.29; p = .005). Neither the effects of presentation mode nor their interaction were significant (p > .50). Consistent with prior studies, the difference in perceived diagnosticity between the attributes was smaller when review volumes were low relative to high (M_{low} = .30, M_{high} = .92). Once again, we find support for mediation (PROCESS Model 4; Preacher, Rucker and Hayes 2007) of the effect of review volumes on preference via the difference in perceived diagnosticity of average product ratings and review volumes (B = -1.46; 95% confidence interval [-.37, -.03]).

Discussion. This study provided further support for the robustness of our effect. Whether consumers evaluate choice options simultaneously or sequentially, the review volumes of considered options play a significant role in the preference between options.

Sample Stimuli
APPENDIX W5

Popularity Cue

The purpose of this study was to rule out product popularity as an explanation for the influence of review volumes on consumer preference. If review volumes were solely cues of popularity, we would expect that another popularity cue (e.g., “Best seller”) would attenuate the influence of review volumes on preference.

Method

Participants and design. We recruited 402 undergraduate students in exchange for course credit. Participants were randomly assigned to a condition in a 2 (review volume levels: low, high) by 2 (popularity cue: absent, present) between-subjects design.

Procedure. We manipulated review volumes across two product options in the same manner as prior studies. In addition, in the popularity cue condition, the higher-rated, fewer reviews option had a “Best Seller” badge. Thus, consumers could choose between an option with fewer reviews but a “Best Seller” badge or an option with more reviews without a badge. Next, participants indicated relative preference between two choice options.

Results. A 2 (review volumes) by 2 (popularity cue) ANOVA on relative preference yielded significant main effects of volume (F(1,398) = 124.43; p < .001) and popularity cue (F(1,398) = 14.90; p < .001). The interaction was not significant (p > .20). Consistent with prior studies, preference for the higher-rated, fewer reviews option was lower in the presence of low versus high review volumes (M_{low} = 4.63 vs. M_{high} = 2.74). Furthermore, preference for the higher-rated, fewer reviews option was also greater when that option was labeled as a “Best Seller” (M_{present} = 3.34 vs. M_{absent} = 4.01).
Discussion. This study ruled out popularity as an alternative explanation. It demonstrated that when consumers are choosing from “Best Sellers”, review volumes still play a critical role in the decision process when those best sellers also have low review volumes. It also did demonstrate a main effect of the presence of the badge such that the mere presence of best-selling options increased preference for the higher-rated one, suggesting that the badge acts as an additional discriminating cue in judgment processes.

Sample Stimuli
Product Versions

The purpose of this study was to demonstrate the persistent effect of review volumes in light of possible justification for low review volumes. In this case, we used product versions (e.g., 2015 vs. 2013 models). Because newer versions have been on the market for less time, their lower review volumes should be justified. Yet, even in this case, we argue that the product version will not attenuate the effect of review volumes, demonstrating that even with proper justification, low review volumes still significantly influence consumer preference.

Method

Participants and design. We recruited 152 participants (M_{age} = 36.43; 52% female; Amazon mTurk; $0.50 payment). Participants were randomly assigned to a condition in a 2 (review volumes: low, high) by 2 (product age information: absent, present) between-subjects by 2 (product replicates: DVD player, phone) within-subject mixed design.

Procedure. We manipulated review volumes across two product options in the same manner as prior studies. In addition, product age information was manipulated slightly differently for each of the product replicated. For the DVD player, we labeled each option with a release year: Option A was a 2015 version while Option B was a 2013 version. For the cell phone, we labeled each option with a version number: Option A was a Galaxy S7 while Option B was a Galaxy S6. Thus, each choice set included a higher-rated, fewer reviews, newer version option and its inverse. For each product replicate, participants indicated relative preference between choice options.

Results. A repeated measures 2 (review volumes) by 2 (product age information) by 2 (product replicate) ANOVA on relative preference yielded significant main effects of review volume (F(1, 148) = 22.59; p < .001) and product age (F(1, 148) = 8.91; p = .003). Neither the
effect of product replicate nor the interactions were significant \( p > .15 \). Consistent with prior studies, preference for the higher-rated, fewer reviews option was lower in the presence of low vs. high review volumes \( (M_{\text{low}} = 3.28 \text{ vs. } M_{\text{high}} = 2.29) \). Furthermore, preference for the higher-rated, fewer reviews option was also greater when it was a newer product \( (M_{\text{new product}} = 2.45 \text{ vs. } M_{\text{no product age information}} = 3.09) \).

Discussion. This study demonstrated the persistent influence of low review volumes even when a low number of reviews is justified by being a newer product. While a main effect existed such that preference for the higher-rated, fewer reviews option was greater when the product’s newer version was disclosed versus when it was not, this disclosure of version did not attenuate the influence of review volumes. As such, this study demonstrated that consumers can be led to prefer older products simply because they also have more reviews. Given the rapid advances in technology, this study provides some evidence that consumers may make suboptimal decisions as a function of review volumes.

Sample Stimuli
APPENDIX W7

New Product Arrival

One reason that a choice may have a higher average product rating and fewer reviews is because it is a newer choice option relative to the competition. As such, it has fewer reviews but better quality (for example, newer technology) leading to a higher average product rating. Next, we test whether introducing new arrival cue will moderate the influence of review volumes on preferences.

Method

Participants and design. We recruited 202 participants from mTurk (M_{age} = 36.37; 43% female) in exchange for a $0.50 payment. Participants were randomly assigned to a condition in a 2 (review volume levels: low, high) x 2 (new arrival cue: absent, present) between-subjects design by 3 (product replicates: tower fan, cookie sheets, knife sets) within-subject, mixed design.

Procedure. We manipulated review volumes across two product options in the same manner as prior studies. In the new arrival cue present condition, Option A also had a badge which denoted that the product was “new”. We used slightly different badges (e.g., “new”, “new arrival”, and “new product”) across the three product categories to improve generalizability. For each product replicate, participants indicated relative preference between the two choice options.

Results. A 2 (review volumes) by 2 (new arrival cue) by 3 (product replicates) repeated-measures ANOVA on relative preference yielded only significant main effect of volume (F(1, 198) = 20.83; p < .001). No other effects were significant (p > .15). Consistent with prior studies, preference for the higher-rated, fewer reviews option was greater in the presence of high review
volumes \((M_{\text{high}} = 2.87)\) relative to when the consideration set featured low review volumes \((M_{\text{low}} = 3.72)\).

**Discussion.** This study demonstrated the robustness of our effect in light of a clear explanation for the low review volume: products being labeled as “new”. Thus, consumers are still likely to demonstrate weakened preference for higher-rated options merely because they have fewer reviews, even when they are newer, a justification for why they would have fewer reviews.

**Sample Stimuli**
Review Credibility

The purpose of this study was to demonstrate the effect of review credibility on how consumers incorporate average product ratings and review volumes into their decisions.

**Method**

**Participants and design.** We recruited 226 undergraduate participants in exchange for course credit. Participants were randomly assigned to one of two review volume levels condition (low, high) between-subjects.

**Procedure.** We manipulated review volumes in the same manner as prior studies. Credibility was manipulated by the presence of a “Consumer Reports Verified” badge for option A, (higher rated, fewer reviews option). After viewing both options, participants were given the choice between Option A, Option B, or to Defer Purchase. Then participants were asked the importance of each attribute.

**Results**

**Option preference.** A 2 (review volume levels: low, high) x 2 (credibility cue: present, absent) binary logit yielded a main effect of review volume levels ($B = -2.90$; Wald $\chi^2 = 7.88$; $p = .005$) on option choice. Neither the effect of credibility nor the interaction were significant ($p > .70$). Consistent with prior studies, preference for the higher-rated, fewer reviews option was greater in the presence of high review volumes ($P_{\text{high}} = 93\%$) relative to when the choice set featured low review volumes ($P_{\text{low}} = 52\%$).

**Choice Deferral.** A 2 (review volume levels: low, high) x 2 (credibility: present, absent) binary logit yielded no significant effects on deferral rates ($p > .20$), though it was directionally
consistent with prior studies where the rate of deferral was greater when review volumes were low ($P_{\text{low}} = .13$) relative to high ($P_{\text{high}} = .05$).

*Meditation via attribute diagnosticity.* As expected, no significant differences existed across image, brand, and price as a function of presentation mode or review volumes ($p > .15$). As before, we calculated the difference in the perceived diagnosticity of average product ratings and review volumes to create a difference measure. A one-way ANOVA on the difference measure yielded a significant effect of review volumes ($F(1, 213) = 11.26; p = .001$). Consistent with prior studies, the difference in perceived diagnosticity between the attributes was smaller when review volumes were low relative to high ($M_{\text{low}} = .50$, $M_{\text{high}} = 1.01$). Once again, we find support for mediation (PROCESS Model 4; Preacher, Rucker and Hayes 2007) of the effect of review volumes on preference via the difference in perceived diagnosticity of average product ratings and review volumes ($B = -.41; 95\%$ confidence interval [-.86, -.14]).

*Discussion.* This study provided further evidence of previous findings while also ruling out review credibility as a potential moderator. Whether the reviews are aggregated from the general population or verified by Consumer Reports, the influence of low review volumes persists. This study did not demonstrate an effect of review volume levels on the rates of deferral. We posit that this could be an effect of the badge as it mitigates the risk from fewer reviews, but not enough to completely shift preference to that option.

Sample Stimuli
APPENDIX W9

ComScore Data Analysis

The purpose of this study was to explore how the retailers’ practice to display review volumes relates to various consumer website interaction metrics commonly available through secondary data. Specifically, we explore whether there is a correlation between the review volume display location (i.e., not displayed, displayed on the search page, displayed on the product details page) and the likelihood of a customer returning to the same retailer during her next shopping session (i.e., website stickiness), the purchase probability (transaction conversion rate or TCR), the number of sessions per visitor, the number of page views per session, and the time spent per session.

Data. Our data comes from ComScore and contains detailed browsing records of over 45,000 households for 12 months period between January 2012 and December 2012 and represents 208,328,731 browsing sessions.

For each session (i.e., defined as a sequence of page requests to the site coming from the same web browser ending with a 30-minute inactivity period), we observe the domain name, the date and time of the site visit, the duration of the visit, and the number of pages requested. We also infer how often consumers choose to transition to a different e-commerce website or come back to the same site (stickiness). We also observe if a purchase was made during each session.

Procedure. We use the observed measures to compute our variables of interest: TCR (probability of purchase), stickiness (self-transitions/all transitions), sessions per visitor (website sessions/website visitors), page views per session (website page views/website sessions), time spent per session (time spent on website/website sessions).
Descriptive Statistics*

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCR</td>
<td>230</td>
<td>.007</td>
<td>.016</td>
<td>.0000</td>
<td>.156</td>
</tr>
<tr>
<td>Stickiness</td>
<td>230</td>
<td>10.48%</td>
<td>10.18%</td>
<td>0.00%</td>
<td>66.87%</td>
</tr>
<tr>
<td>Number of Sessions per Visitor</td>
<td>230</td>
<td>2.50</td>
<td>2.61</td>
<td>1.00</td>
<td>23.99</td>
</tr>
<tr>
<td>Number of Page Views per Session</td>
<td>230</td>
<td>6.94</td>
<td>4.06</td>
<td>1.00</td>
<td>29.00</td>
</tr>
<tr>
<td>Time Spent per Session</td>
<td>230</td>
<td>5.31</td>
<td>3.052</td>
<td>.00</td>
<td>20.00</td>
</tr>
</tbody>
</table>

*We limit our analyses to the websites in the three high-grossing product categories (home electrics, electronics, apparel; for detail, see footnote 1 in the paper), resulting in 4,072,999 unique sessions for 230 websites.

Results

Transaction Conversion Rate (TCR). A one-way ANOVA of review volumes display on TCR exhibits a significant effect (F(2,227) = 10.32; p < .001). Planned contrasts demonstrate that the retailers who display review volumes on the product page have a higher TCR than those who do not display review volumes (M_{absent} = .001, M_{product} = .006; t(227) = 2.16; p = .032). Furthermore, retailers who display review volumes on the search page have a higher TCR than those who display them on the product page (M_{search} = .013; t(227) = -2.65; p = .009).

Stickiness. A one-way ANOVA of review volumes display on stickiness exhibits a significant effect (F(2,227) = 18.87; p < .001). Planned contrasts demonstrate that the retailers who display review volumes on the product page have a higher stickiness than those who do not display review volumes (M_{absent} = 5.21%, M_{product} = 11.35%; t(227) = 4.10; p = .001). Furthermore, retailers who display review volumes on the search page have a higher stickiness than those who display them on the product page (M_{search} = 14.89%; t(227) = -2.32; p = .021).

Number of sessions per visitor. A one-way ANOVA of review volumes display on sessions per visitor exhibits a significant effect (F(2,227) = 8.36; p < .001). Planned contrasts demonstrate that the retailers who do not display review volumes have fewer sessions per visitor than those who display review volumes on the search page (M_{absent} = 1.51, M_{search} = 3.13; t(227)
= 3.80; \( p < .001 \) and the product page (\( M_{\text{product}} = 2.82; t(227) = 3.27; p = .001 \)). There is no significant difference in sessions per visitor across search and product page displays (\( p > .40 \)).

**Number of page views per session.** A one-way ANOVA of review volumes display on sessions per visitor exhibits a significant effect (\( F(2,227) = 9.68; p < .001 \)). Planned contrasts demonstrate that the retailers who do not display review volumes have fewer page views per session than those who display review volumes on the search page (\( M_{\text{absent}} = 5.34, M_{\text{search}} = 8.13; t(227) = 4.23; p < .001 \)) and the product page (\( M_{\text{product}} = 7.33; t(227) = 3.23; p = .001 \)). There is no significant difference in page views per session across search and product page displays (\( p > .20 \)).

**Time spent per session.** A one-way ANOVA of review volumes display on sessions per visitor exhibits a significant effect (\( F(2,227) = 8.50; p < .001 \)). Planned contrasts demonstrate that the retailers who do not display review volumes have less time spent per session than those who display review volumes on the search page (\( M_{\text{absent}} = 4.20, M_{\text{search}} = 6.22; t(227) = 4.04; p < .001 \)) and the product page (\( M_{\text{product}} = 5.51; t(227) = 2.80; p = .006 \)). There is no significant difference in time spent per session across search and product page displays (\( p > .10 \)).

**Means by Review Volumes Display**

<table>
<thead>
<tr>
<th></th>
<th>Not Visible</th>
<th>Visible on Search page</th>
<th>Visible on Product Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCR</td>
<td>.001</td>
<td>.013</td>
<td>.006</td>
</tr>
<tr>
<td>Stickiness</td>
<td>5.21%</td>
<td>14.89%</td>
<td>11.35%</td>
</tr>
<tr>
<td>Number of Sessions per Visitor</td>
<td>1.51</td>
<td>3.13</td>
<td>2.82</td>
</tr>
<tr>
<td>Number of Page Views per Session</td>
<td>5.34</td>
<td>8.13</td>
<td>7.33</td>
</tr>
<tr>
<td>Time Spent per Session</td>
<td>4.20</td>
<td>6.22</td>
<td>5.51</td>
</tr>
</tbody>
</table>

**Discussion.** Exploring site browsing data from ComScore, we find that the retailer decision of whether and where to display review volumes shows significant correlation with
various consumer website interaction metrics. One might argue that the data from larger retailers (e.g., Amazon.com) could be overly influencing these results as they often have similar strategies to display product reviews, and also have many more sessions than smaller retailers. To address this concern, we also controlled for the number of sessions for each retailer (an approximation of their popularity), and found that the same patterns emerge and are significant (all $p’s < .002$), suggesting that it is not merely a site traffic effect.

We must also note that, unfortunately, our dataset does not include detail record of what products were considered but not purchased in each session (i.e. choice set information) to make a more direct connection of this analysis to the theory presented and tested in the paper. Also, the presented results should be interpreted as correlational without any causality implied. Nevertheless, we find the data patterns revealed to be interesting and consistent with the argument presented in the paper that the review display format have an impact on consumer onsite behavior.