

Author Accepted Manuscript



**Quality in Context: Experience-Relevant Consumption
Context Influences Product Ratings**

Journal:	<i>Journal of Marketing</i>
Manuscript ID	JM-24-0608.R2
Manuscript Type:	Revised Submission
Research Topics:	Customer Reviews, Digital Marketing - Customer, Online Reviews, Online Retailing, User-Generated Content, Word of Mouth
Methods:	Lab Experiments, Multimethod, Online Experiments, Regression Models

SCHOLARONE™
Manuscripts

Quality in Context: Experience-Relevant Consumption Context Influences Product Ratings

Matt Meister

Assistant Professor of Marketing, University of San Francisco

2130 Fulton Street, San Francisco, CA 94117-1080

mmeister@usfca.edu

415-422-6721

Nicholas Reinholtz

Assistant Professor of Marketing, University of Colorado

995 Regent Drive, Boulder, CO 80309-0419

nicholas.s.reinholtz@colorado.edu

303-735-8019

Matt Meister is an Assistant Professor of Marketing at the University of San Francisco. Nicholas Reinholtz is an Assistant Professor at the University of Colorado Boulder Leeds School of Business. Communication should be directed to Matt: mmeister@usfca.edu. The authors thank Quentin André, Stephen A. Spiller, John G. Lynch, Phillip M. Fernbach, Ryan Lewis, Leaf Van Boven, and seminar participants at the University of California at Berkeley, University of Toronto, Ivey Business School at Western University, University of Colorado Boulder, and Colorado Winter Conference on Marketing and Cognition for helpful comments. The authors have no financial conflicts to report.

Quality in Context: Experience-Relevant Consumption Context Influences Product Ratings**Abstract**

Experience with a product is shaped by two things: (i) aspects of the product itself and (ii) aspects of the environment in which the product is consumed. This paper documents evidence that when translating their experiences into ratings for products, consumers overly attribute their experience to products, and under-attribute experience to context—the environment in which the product is consumed. First, 218,918 ratings collected from REI.com demonstrate that ratings for cold-weather gear (products designed to keep people warm) are positively correlated with temperature: These products get lower ratings when the weather is cold and higher ratings when the weather is warm, controlling for climate and season. Ratings for other products (e.g., bicycles, tents, skis) are not affected by temperature. This effect generalizes to other products and contexts both in the REI data (e.g., rain jackets and rain) and in a laboratory experiment. The paper also identifies attenuating conditions for this effect. Specifically, the effect seems to be smaller when reviewers explicitly consider context, and when context information is made more accessible, while rating. These findings inform several possible interventions for platforms, which are assessed and validated.

Keywords: User-generated content, online reviews, context, weather.

Introduction

User-generated ratings and reviews are a common feature of the online consumer experience. After consumption experiences, consumers are asked to evaluate products (e.g., Amazon.com), restaurants (e.g., Yelp.com), lodging (e.g., Airbnb.com), and even medical doctors (e.g., Healthgrades.com). These ratings are then aggregated and shown to prospective consumers, who use them to make purchase decisions, specifically treating ratings as a reliable way to distinguish the quality of alternatives (Chen, Wang, and Xie 2011; Chintagunta, Gopinath, and Venkataraman 2010; Dellarocas, Zhang, and Awad 2007). Some have argued that user-generated ratings are a boon for prospective consumers, as they summarize the experience of past consumers (Simonson and Rosen 2014). In theory, this allows someone to get a sense for what their experience would be like if they choose the same option. In this paper, we suggest this benefit may also come with a cost: The experience of past consumers may mislead future consumers if they do not share the same consumption context.

A consumer’s experience with a product is shaped by an interaction of two things: (i) features intrinsic to the product itself and (ii) contextual factors that are extrinsic to the product. For example, a consumer’s experience with a winter jacket is shaped by the jacket’s style, build quality, technical features like extra pockets, type and amount of insulation, and so on. These intrinsic features are shared by all consumers of the same winter jacket. Meanwhile, a consumer’s experience with a winter jacket is also shaped by the weather: temperature, wind, rain, and so on. These extrinsic factors are a property of the context in which the jacket is consumed and may differ drastically across consumers who experience the same winter jacket.

Author Accepted Manuscript

The usefulness of user-generated ratings depends on their ability to convey aspects of the consumption experience shared by past and prospective consumers. This could be accomplished if consumers rate products based solely on their intrinsic features, which—by definition—are shared by future consumers. Potentially, this could also be accomplished if raters factor in extrinsic inputs to their consumption experience, and if prospective consumers are aware of these contextual factors and can adjust their forecasted experience accordingly.

Unfortunately, our results contradict both of these possibilities: We find that ratings are systematically influenced by experience-relevant consumption context in ways that are effectively invisible to prospective consumers. For instance, people give lower ratings to the same winter jacket when they wear that jacket in colder weather compared to warmer weather. Meanwhile, the raters who mention context in their reviews—and thus are more likely aware of its influence on their experience—do not show the effect. Moreover, raters are more consistent when prompted to consider their consumption context prior to providing a rating. These results demonstrate that consumption context currently poses a threat to the usefulness of ratings.

The uncorrected effect of consumption context on ratings can lead to issues of both bias and noise. Bias can emerge when two otherwise similar products are consumed in systematically different contexts. Noise emerges when consumption context is effectively random across products, and is problematic because ratings distributions with high variance increase consumers' uncertainty about products, decreasing choice (He and Bond 2015; Matz and Wood 2005; Urbany, Dickinson, and Wilkie 1989). Additionally, noise is particularly problematic in cases where there are only a small number of ratings (i.e., the law of large numbers does not apply), where increased noise impacts differences in product rankings when sorted by ratings. As

we later show, the context-driven effect on ratings we observe can lead to fairly drastic changes in the rank order of products when sorting by average rating, which prior research suggests will have substantial downstream effects on consumer search (Ursu 2018).

Fortunately, there are things platforms can do to mitigate the effect of context on ratings. In situations where the context is observable to the platform, they can attempt to correct for it algorithmically. We provide a simple demonstration of how this might work in the case of temperature in the discussion of Study 1. In situations where context is not observable, our results suggest that platforms can improve their collection of ratings by using prompts which encourage consumers to consider the influence of context on their consumption experience.

The rest of the paper is organized as follows: First, we discuss past research on consumer (mis)attribution and how it relates to our hypotheses and the present investigation. Next, we provide an overview of empirical evidence. After this, we present the four studies individually. We conclude with a general discussion, focusing on downstream consequences and elaborating on the aforementioned practical implications for manufacturers and platforms.

Conceptual Development

Consumer (Mis)attribution

When using ratings, consumers act as if they consider ratings to be a measure of products' intrinsic quality. For example, they are more likely to select higher rated products (Chen et al. 2011; Chintagunta et al. 2010; Dellarocas et al. 2007) and expect those higher rated options to provide more utility (de Langhe, Fernbach, and Lichtenstein 2016). This implies that ratings are most useful to consumers when they convey a product's quality.

Author Accepted Manuscript

Rating a product in terms of its quality presents a challenge of attribution. This is because experience with a product is influenced by both the product itself and the environment in which the product is consumed. If the consumer had a positive experience with a product, it could be because the product is great (e.g., a high quality jacket), because they consumed it in a favorable environment (e.g., warmer weather), or both. To translate their experience into a rating that reflects the product's quality, the consumer must identify—and effectively control for—the relative influence that consumption context had on their experience.

A long history of research in psychology, marketing, and economics suggests that humans are not particularly good at this type of attribution task. A prominent example comes from the domain of social cognition. In a typical study, participants read about the behavior of an actor in a certain situation. Then, participants are asked to judge the degree to which they think that behavior should be attributed to properties of the actor (e.g., their beliefs or preferences) or properties of the context/situation. Findings suggest that people neglect contextual influences and attribute behavior to actors. In a classic example, participants judged a student who was assigned to write a pro-Castro essay to have personally endorsed Castro's views (Jones and Harris 1967). This misattribution of outcomes to actors instead of context is so robust and pernicious, it has been called the fundamental attribution error (Gilbert and Malone 1995; Ross 1977).

Another well-documented form of misattribution involves the role of incidental affective states on judgments and evaluations. A key idea in this work is that people use affect as a source of information for complex judgments (i.e., people assume that positive emotion is caused by the thing they are evaluating being good; Cohen, Pham, and Andrade 2018; Schwarz and Clore 1983). However, misattribution occurs when affect is unrelated to the focal evaluation. In a classic example, men who encountered a woman after crossing a precarious, wobbly bridge

Author Accepted Manuscript

seemed to misattribute their physiological arousal to romantic attraction, not realizing their racing hearts were actually caused by the situational context (Dutton and Aron 1974).

The misattribution of incidental affect (“affect-as-information”) is particularly relevant for the current investigation, as many previous demonstrations involve consumer products and evaluations. For example, Imschloss and Kuehnl (2019) conclude that consumers evaluate fabrics as softer when stores play soft ambient music, because they misattribute their emotional reaction to the music to the fabric. Similarly, Meyers-Levy, Zhu, and Jiang (2009) find consumers evaluate products more negatively when they stand on uncomfortable flooring, presumably because they misattribute feelings of discomfort to the products in question.

Most closely related to our investigation, Brandes and Dover (2022) show that consumers ratings for past hotels stays are influenced by the weather they are currently experiencing: When asked to rate a past stay during poor (cold, rainy) weather at home, they tend to give that hotel lower ratings than when rating the past day during good (warm, sunny) weather. They suggest this is because the weather they are currently experiencing affects their current mood (consistent with Schwarz and Clore 1983), which they misattribute to their past hotel experience.

Misattribution of Experience-Relevant Context

In contrast to examples in which people broadly misattribute incidental affect to unrelated experiences (e.g., mood today might affect ratings of a hotel stay from the past), our hypotheses regard contextual factors that influence the actual consumption experience. For example, experience with a winter jacket depends on the temperature outside; experience with a rain jacket depends on precipitation; the experience of consuming an energy drink depends on how tired one feels before consumption. Context can have positive or negative effects on experience: In favorable contexts the experience will be better (e.g., warmer weather for a jacket) and in

Author Accepted Manuscript

unfavorable contexts the experience will be worse (e.g., colder weather for a winter jacket). We propose that this effect on experience is likely to be misattributed to the product being evaluated.

The misattribution we predict is specific to contexts which are directly relevant to the experience of specific products. A key difference between our experience-based account and the mood-based account in prior work is the scope of the effect: Whereas the mood-based account should apply broadly to all product categories, the experience-based account we propose should apply differently to product categories for which a specific form of context strongly affects experience. Therefore, we distinguish our experience-based account directly by comparing the effect of context across products for which that context is and is not directly experience-relevant. Specifically, we predict that experience-relevant context has a significant impact on ratings:

H₁: Experience-relevant context will have a significant influence on ratings, such that contexts which are favorable to experience will lead to higher ratings than contexts which are unfavorable to experience.

We note as well that ratings are influenced by many factors, such that the impact of experience-relevant context is not the only factor that causes ratings to deviate from objective quality. Even if the effect of context was removed entirely, user generated ratings would still remain far from perfect. For a recent example, heightened expectations have been shown to bias ratings downwards; consistent with the broader expectation-disconfirmation literature (Oliver 1977), Meister and Reinholtz (2025) find that the same Airbnb properties earn lower ratings on average when they are listed as “Superhosts” (v. regular hosts), because this status raises consumers’ *a priori* expectations without altering quality. Prior consumers’ ratings have also been shown to have a social influence on subsequent ratings, lessening the impact of positive and negative product attributes (Sridhar and Srinivasan 2012). For example, Park, Shin, and Xie

(2021) find that the valence of a product’s first rating has a measurable impact on all subsequent ratings. Memory is also likely to play a role, as consumers are unlikely to create ratings with all aspects of an experience top of mind. Thus, experience-relevant context is not the only reason that ratings stray from objectivity. However, its impact has thus far been overlooked in marketing research, as have ways to mitigate this effect, which we propose in this paper.

Mitigating Role of Awareness

We propose that the influence of experience-relevant context on ratings will be diminished when the influence of context on experience is salient to the rater. This is consistent with other cases of misattribution, where making relevant contributing factors salient reduces the extent of misattributions. In the previously cited example of correspondence bias—where participants judged the political views of essay writers (Jones and Harris 1967)—participants judged writers as significantly less pro-Castro when informed that these writers were instructed to write pro-Castro essays. The related literature on affect-as-information shows similar attenuation with awareness. For instance, Schwarz and Clore (1983) found that participants reported lower life satisfaction on rainy days than on sunny days. They argue this pattern is because weather influenced moods, which participants misattributed to be caused by their overall life satisfaction. However, Schwarz and Clore found no difference in satisfaction when participants were reminded of weather. Their argument is that context clarified for participants the actual cause of their mood, leading them not to misattribute it to life satisfaction.

Because increased awareness of contributing factors is a common mechanism across misattribution research broadly, we predict that awareness of experience-relevant consumption context will also mitigate the extent to which that context influences consumers’ ratings. We

Author Accepted Manuscript

predict that when consumers are aware of their context—such that they can attribute their product experience to it—context will have less of an impact on their ratings. Formally:

H₂: The effect of experience-relevant consumption context on user-generated ratings will be attenuated by users' awareness of their context.

Empirical Overview

We present the results of four studies in this paper, with two further in the web appendix. Study 1 is an analysis of product reviews collected from REI.com (and outdoor equipment retailer) and merged with weather data. We find that ratings for cold-weather products (e.g., winter jackets) systematically differ depending on temperature: When it is colder outside, cold-weather products receive lower ratings, presumably because the raters feel cold and misattribute this feeling to products. Supporting this experience-relevant misattribution account, the effect of temperature on ratings is significantly smaller for other, non-cold-weather products (e.g., bicycles), for which keeping the user warm is not the main purpose. Importantly, our preferred identification strategy controls for variation in temperature and ratings caused by local seasonality, and differences in temperature and quality between different products, giving credence to a causal interpretation of our results. Exploratory text analyses suggest consumers who mention context in reviews—and thus for whom the temperature while they were consuming the product was likely salient while rating—provide ratings that are less influenced by context. Study 2 further explores the REI ratings, showing that a similar effect emerges in another product category and context—rain jackets and rain.

Studies 3 and 4 extend these findings in an experimental context, with a focus on establishing the underlying process and assessing possible mitigation strategies for platforms. Study 3 documents the effect across a variety of different products and consumption contexts,

and provides further evidence that the salience of context reduces its effect on ratings. Finally, Study 4 assesses four different strategies (informed by the results from the prior studies) that platforms could use to improve their collection of ratings. Results suggest that platforms can mitigate effects of context by overtly reminding raters to consider the context of consumption while creating their rating, but only the most specific reminders resulted in substantial attenuation effects. Data, code, and materials (including preregistrations for Studies 3 and 4) are available on OSF (https://osf.io/vbax7/?view_only=f495664d832344aebd49b1ec549321e0).

Study 1: Consumption Context Affects Ratings on REI.com

In this study, we explore how temperature at the time of consumption influences ratings of products on REI.com. We distinguish between two possible mechanisms for a relationship between temperature and ratings: (i) an affect-based account in which temperature influences ratings for all products and (ii) our consumption-context account that predicts temperature will only influence ratings for products whose purpose is to keep people warm.

We first describe our data. Next we describe our general approach to empirical model construction. Then, we present the results of our preferred specification (Model 5) in Table 1, alongside a series of less-constrained models (Models 1–4) with fewer controls relative to the preferred, comprehensive specification. Finally, we present a specification curve analysis, which assesses the robustness of the results across alternative model choices.

Data

We collected all available user-generated reviews from REI.com by first collecting all product links and then retrieving the maximum possible number of reviews for each product.¹

¹ The platform only returned the most recent 100 reviews for each product during each wave of data collection. Because we ran three waves of data collection (in February and October 2022, and November 2023; duplicate reviews were removed), the maximum number of reviews for a product in our data is 300.

Author Accepted Manuscript

This process left us with 558,365 ratings, of which 461,821 were accompanied by a text review. For each rating, we collected all information available from REI.com, including the text of the review, reviewer ID ($N = 477,015$), product ID ($N = 20,357$), product category ($N = 186$), rating (1–5), and reviewer location ($N = 275,396$, free response). Consistent with past research, average ratings are fairly high (mean of average ratings = 4.26/5, median = 4.44; cf Schoenmueller, Netzer, and Stahl 2020; Zervas, Proserpio, and Byers 2021) and most products have a small number of ratings (mean number of ratings = 27.43, median = 0; cf de Langhe et al. 2016).²

REI.com assigns products to different categories. We identified ten categories that contain cold-weather gear (products intended to keep the user warm): gloves and mittens, snowboard clothing, ski clothing, men's and women's jackets, men's and women's snow jackets, and men's and women's insulated jackets. Our prediction was that ratings for products listed in these categories would particularly depend in part on the temperature at the time they were used, as temperature represents context that would affect the consumer's experience of these products.

We supplemented the REI.com data with daily weather observations from the National Centers for Environmental Information (NCEI). To do this, we use the location and date information in the review data. We began with the 275,396 reviews which included locations. Because this is captured by a free response field, there is no standard format, nor latitude and longitude provided by REI. Instead, we parsed location information where we could from common formats such as city, state name or city, state abbreviation. We then merged these with a list of 28,338 US cities, which included latitude and longitude coordinates. When we could

² We note two ways in which our data set deviates from the total population of products and ratings/reviews on REI.com: (i) we do not include products with zero ratings and (ii) we have (at most) 300 ratings for each product. Using data captured on the product pages, we can calculate the median (6) and mean (59.52) number of reviews for all products on REI.com at the time of data collection.

only parse a user’s state, we assumed they were from the population center of that state.³ We then calculated the closest NCEI weather station ($N = 2,148$) for each review and merged the reviews and weather data on review date and this station. This process yielded our final sample of 218,913 reviews for which we could attach weather data. Of these, 20,060 (9.16%) were for cold-weather gear, and 198,853 (90.84%) were for other products.

A limitation of our data is that we do not know exactly when a product was consumed, only when the review was submitted. So, in our analyses we use a measure of the local temperature over a window of days prior to the date of the review as a proxy for the temperature at the time of consumption. In our focal specification, we use a 3-day average (the day of the review and the two days prior).⁴ In further analyses, we show the results we report are robust to other operationalizations of consumption temperature (i.e., using minimum or maximum temperature instead of an average; using shorter or longer time windows).

Main Analysis

To assess the role of consumption context on product ratings, we examine the degree to which temperature at the time of consumption influences ratings for products intended to keep the consumer warm. We predict that ratings for cold-weather gear are positively correlated with temperature, and that temperature has a greater effect on ratings for cold-weather gear than for other products (for which warmth is not a main purpose).

We test this prediction in a series of regression models in which an individual rating (discrete: 1–5) serves at the unit of analysis. In each model, we include two focal variables and their interaction. First, temperature (in degrees Fahrenheit), which in our preferred specification

³ Results are robust to dropping these observations, as we do in the specification curve analysis reported later.

⁴ See Web Appendix A for ratings histograms and temperature density plots.

Author Accepted Manuscript

is the three-day average at the rater's location. Second, we include an indicator variable for whether the rating is for cold-weather gear (1 = cold-weather gear, 0 = not). Finally, we include an interaction between temperature and the cold-weather gear indicator variable. This coefficient reflects the differential effect of temperature on ratings for cold-weather gear compared to other types of products, and thus reflects the focal test for H_1 : A positive interaction coefficient suggests that temperature affects ratings for cold-weather gear more than for other products.

We present the results of a series of regression models building up to our preferred specification (Model 5) in Table 1. Model 1 includes only the three predictor variables described above. This model does not control for local climate, seasonality, or differences between products. Latter models add progressively granular fixed effects for location, month, and product, removing variation caused by those factors. Model 2 adds weather station fixed effects, which remove variation in ratings and temperature across locations. Model 3 adds month fixed effects, which remove variation caused by seasonality in both ratings and temperature. Model 4 includes both weather station and month fixed-effects, which remove variation across locations *and* seasons. These fixed effects effectively control for "normal" temperature in the same location and month as the observed review. Finally, Model 5 (our preferred specification) additionally includes product fixed effects, which control for any remaining systematic differences across products. In all models, standard errors are robust and clustered by product.

Because of the different fixed-effect structures, the effect of interest is estimated on different variation in temperature across the different models. In Model 1, the effect is estimated on all variation in temperature, including seasonal changes and differences across locations. We anticipate that these differences in temperature would affect experience with cold-weather gear

and thus lead to differences in ratings (H_1). However, the systematic differences in temperature between seasons and locations confound the effect of temperature with season and geography.

Model 5 addresses the confounds. The effect of temperature in this model is estimated only on “abnormal” temperature—deviations in temperature from the seasonal average at the specific location. These deviations are plausibly exogenous (Dell, Jones, and Olken 2014), and thus the temperature effect in Model 5 is most amenable to a causal interpretation.

Main Results

Results from all models support our prediction that experience-relevant consumption impacts ratings: The interaction between product type and temperature is positive across all specifications. This indicates that ratings for cold-weather products vary depending on temperature (e.g., Table 1, Model 5; $\beta_{Interaction} = .002$, $t(199,725) = 3.93$, $p < .001$, 95% CI = [.0011, .0034], Std. $\beta = .032$) to a greater degree than ratings for non-cold-weather gear.

Taking a causal interpretation of the results from Model 5 suggests that differences in abnormal temperature do not affect ratings for non-cold-weather gear (e.g., $\beta_{Temp} > -.001$, $t(199,725) = -1.327$, $p = .185$, 95% CI = [-.0012, .0002], Std. $\beta = -.007$). But, that differences in abnormal temperature *do* cause differences in ratings for cold-weather gear (re-coding the CWG indicator such that it is zero for cold-weather gear: $\beta_{Temp} = .002$, $t(199,725) = 2.982$, $p = .003$, 95% CI = [.0006, .0039], Std. $\beta = .025$).

Table 1: Regression Results for Main Analyses in Study 1.

	Dependent Variable: Rating (1-5)				
	(1)	(2)	(3)	(4)	(5)
Temp	-0.001 (.0002)	-0.002 (.0003)	0.002 (.0003)	-0.001 (.0004)	-0.0005 (.0004)
CWG	-0.074 (.038)	-0.091 (.038)	-0.037 (.039)	-0.07 (.038)	- -

Author Accepted Manuscript

Temp x CWG	0.004 (.001)	0.004 (.001)	0.003 (.001)	0.004 (.001)	0.002 (.001)
Constant	4.269 (.014)	- -	- -	- -	- -
Station FE	No	Yes	No	Yes	Yes
Month FE	No	No	Yes	Yes	Yes
Product FE	No	No	No	No	Yes
Observations	218,913	218,913	218,913	218,913	218,913
R ²	0.001	0.018	0.002	0.018	0.208
Adjusted R ²	0.001	0.008	0.002	0.009	0.132
Residual Std. Error	1.242	1.237	1.241	1.237	1.157
df	218,909	216,762	218,898	216,751	199,725

Note: “Temp” refers to the effect of one degree increase in the average daily mean temperature in the day of a review and two prior. “CWG” is an indicator that equals 1 if the product is cold-weather gear. In all models, SEs are robust and clustered by product.

Robustness

Specification curve analysis

The models described in the previous section reflect principled, but specific choices about construct operationalizations (e.g., a three-day average for temperature) and model structure (e.g., choice of fixed effects). We assess the robustness of the main result to alternative choices using a specification curve analysis (Simonsohn, Simmons, and Nelson 2020). In this analysis, we identify elements of our model that could change (e.g., the temperature measure) and different possibilities for each of these elements (e.g., average, minimum, or maximum). We estimate the interaction of temperature and the cold-weather gear indicator in every combination of these different choices, yielding 20,736 unique specifications.

Table 2: Each Analytic Choice in Study 1 Specification Curve.

Element	Unique Choices	Specific Choices
Data	2	All/No Imputed Locations
Temp Measure	3	Mean/Min/Max
Temp Horizon	3	1/3/7 Day
Abnormal Measure	3	Raw/Minus Recent/Minus Normal
Primary FEs	4	None/Station/Station×Month/Station×Week
Other FEs	8	None/Product/Brand/Year
Controls	12	None/Price/Price (Winsorized)/N Reviews/Precipitation

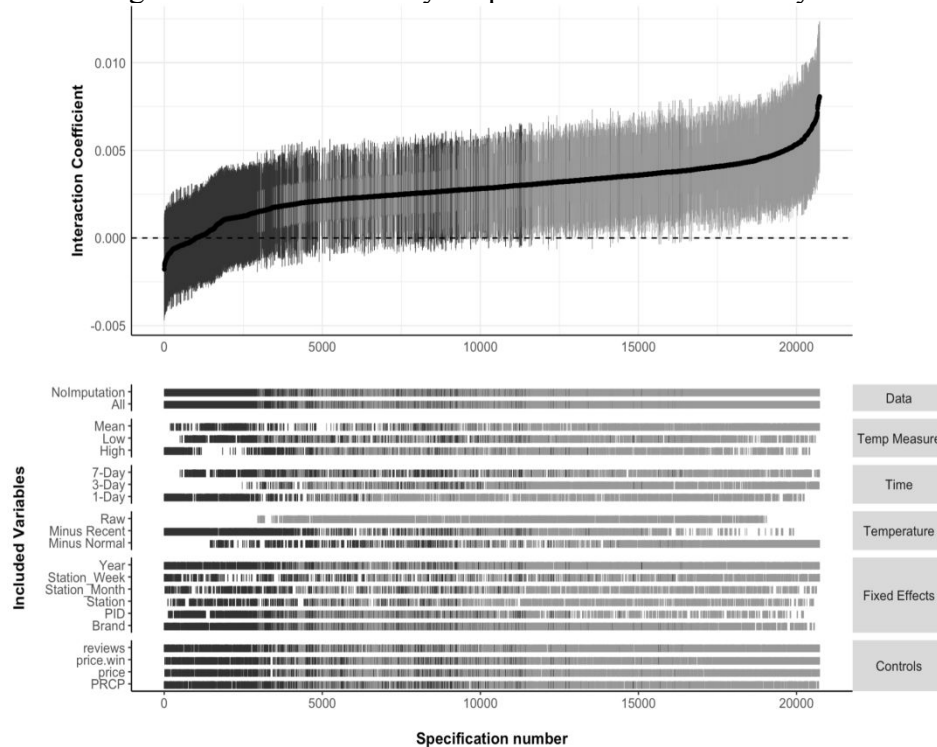
The different elements and choices are summarized in Table 2. The analysis varies whether we use the entire data set (including ratings for which the location is just the state), or only the subset of ratings with city-specific location information. Each specification uses one of 27 ($3 \times 3 \times 3$) operations of temperature, varying the measure (mean, minimum, or maximum), horizon over which it is calculated (1, 3, or 7 days), and whether we use the raw temperature or subtract the most recent or the historical average over the same time period from the raw temperature. We vary fixed-effect structures, including fixed effects that capture location-by-season differences, product differences (product and brand) and long-run changes in rating behavior (year). Finally, we vary the inclusion of different possible control variables. As in Models 1–5, all models use robust standard errors, clustered by product.

Results from the specification curve analysis are shown in Figure 1. The coefficient estimate from each model for the temperature-by-cold-weather-gear interaction is plotted as a black dot and the 95% confidence interval for each estimate is plotted as a grey line. Results are ordered (left-to-right) by the magnitude of the coefficient estimate.

The main result is robust across specifications: The interaction coefficient was positive 19,738 (95.2%) of the possible specifications and the 95% confidence interval excluded zero in 14,926 (72.0%). Inspection of the coefficients across different models reveals a few systematic differences that arise based on specification. The smallest estimates all feature a transformed version of temperature (raw minus recent or raw minus historical) and use 1- or 7-day windows.

Author Accepted Manuscript

Figure 1: Results of Study 1 Specification Curve Analysis.



Note: Vertical ticks in the bottom panel indicate the different choices made for each model.

Individual categories

To investigate whether the effect generalizes to all cold-weather gear, we replicated the regression of Table 1, Model 5 on each of the ten categories that make up cold-weather gear (e.g., gloves and mittens, men's jackets). In each, we consider only one category of cold-weather products at a time. Table 3 presents the interaction coefficient of product type and temperature from these replications of Table 1, Model 5 within individual categories. Results presented support the generalizability of our findings: The point estimate is positive for every category but one—snow clothing, which has only 47 observations.⁵

Table 3: Regression Results Within Cold-Weather Categories

⁵ The smaller estimate for snowboard clothing is consistent with the notion that our proxy measures for location and date reduce our effect size. Because relatively few people live at places where they can snowboard, this category is likely to have many reviews for consumption experiences that took place far from reviewing locations.

Category	Ratings	Products	Station:Months	Int. Est.	p
Men's Insulated Jackets	1,615	80	956	0.005	0.002
Men's Snow Jackets	1,213	99	770	0.005	0.078
Gloves and Mittens	1,937	132	958	0.004	0.065
Ski Clothing	561	103	381	0.003	0.281
Women's Snow Jackets	2,125	127	1,097	0.003	0.083
Women's Insulated Jackets	1,826	63	1,029	0.002	0.155
Men's Jackets	3,121	305	1,611	0.002	0.241
Women's Jackets	2,557	293	1,321	0.001	0.419
Snowboard Clothing	5,058	542	1,926	0.001	0.244
Snow clothing	47	11	41	-0.026	0.007

Above v. below average temperatures

Abnormal differences in temperature can be either warmer than normal or colder than normal. To assess whether the effect we observe is similar across the range of temperatures, we augmented our preferred model (Table 1, Model 5) by interacting an indicator for “above normal” (1 if above the historical average) with raw temperature, product type, and their interaction. Results suggest a stronger effect for temperatures above normal: The interaction of the above normal indicator, product type, and temperature ($\beta = .0022$, $t(199,704) = 2.101$, $p = .036$, 95% CI = [$< .0001$, $.0004$], Std. $\beta = .031$) indicated that the effect of consumption context was more positive for above normal temperatures ($\beta = .0036$, $t(199,704) = 4.142$, $p < .001$, 95% CI = [$.0002$, $.0053$], Std. $\beta = .051$) than for below normal temperatures ($\beta = .0014$, $t(199,704) = 1.923$, $p = .055$, 95% CI = [$> -.0001$, $.0003$], Std. $\beta = .020$).⁶

Sentiment: Analysis & Results

Consumers read text reviews alongside ratings (Bambauer-Sachse and Mangold 2011; Varga and Albuquerque 2019), and reviews provide richer information than ratings alone (Tirunillai and Tellis 2014). In total, we observe 218,900 text reviews (13 ratings had location information but no text). We analyze these reviews for two reasons; first, to replicate our main

⁶ We also binned each observation according to deciles of abnormal temperatures. Figure A1 in the appendix presents the mean and 95% confidence interval of ratings for products within these bins, separated for cold-weather and other products. This shows the same result.

Author Accepted Manuscript

analyses with review sentiment, and second to investigate differences between reviewers who mention context—indicating awareness of that context—and those who do not.

Sentiment analysis

First, we replicate the analyses in Table 1 using sentiment. Sentiment was measured using the large language model *GPT-4o-mini* (Achiam et al., 2023), an approach that has been previously validated (e.g., Abdurahman et al. 2024; Belal, She, and Wong 2023; Rathje et al. 2024). We chose to use GPT due to its ability to measure sentiment more specifically in-context than simpler methods. We sent GPT reviews individually, with the prompt “*I will give you a product review. Provide one response. Is the text of this review more negative (0), or positive (9) towards the product? Respond with only a single integer (0-9).*” Due to cost considerations, we did not submit non-cold-weather gear reviews from location-month combinations that had fewer than four cold-weather reviews. We submitted all other reviews; 20,060 for cold-weather gear, and 111,014 for non-cold-weather gear.⁷

Table 4: Regression Results with Sentiment as the Dependent Variable.

	<i>Dependent Variable: Sentiment</i>				
	(1)	(2)	(3)	(4)	(5)
Temp	-0.004 (.001)	-0.004 (.001)	-0.002 (.001)	-0.001 (.001)	-0.0002 (.001)
CWG	-0.261 (.097)	-0.311 (.097)	-0.196 (.097)	-0.258 (.102)	- -
Temp × CWG	0.012 (.002)	0.011 (.002)	0.01 (.002)	0.01 (.002)	0.007 (.002)
Constant	7.106 (.039)	- -	- -	- -	- -
Station FE	No	Yes	No	Yes	Yes
Month FE	No	No	Yes	Yes	Yes

⁷ In web appendix B, we present a “bag of words” analysis of review text, which yields similar results.

Product FE	No	No	No	No	Yes
Observations	131,064	131,064	131,064	131,064	131,064
R ²	0.002	0.014	0.003	0.036	0.257
Adjusted R ²	0.002	0.006	0.003	0.003	0.125
Residual SE	3.04	3.034	3.038	3.038	2.847
df	131,060	129,925	131,049	126,765	111,267

Note: “Temp” refers to the effect of one degree increase in the average daily mean temperature on the day of a review and two prior. “CWG” is an indicator that equals 1 if the product is cold-weather gear.

Most reviews were coded as either fully positive (i.e., 9; 56.3%) or fully negative (i.e., 0; 10.4%), with the remainder roughly uniform across the medial values. The correlation between rating and sentiment was quite high ($r = .894$). Thus, it is unlikely that the text of reviews contains much information about overall sentiment beyond what is expressed by ratings. Table 4 presents estimates from five models, which replicate Table 1 with sentiment as the outcome.

Results demonstrate a consistent positive interaction effect of temperature on sentiment in the first five models (e.g., Model 5; $\beta_{Interaction} = .007$, $t(111,267) = 4.205$, $p < .001$, 95% CI = [.004, .011], Std. $\beta = .042$): People write more negative reviews for cold-weather gear—but not other types of products—after periods of colder weather.

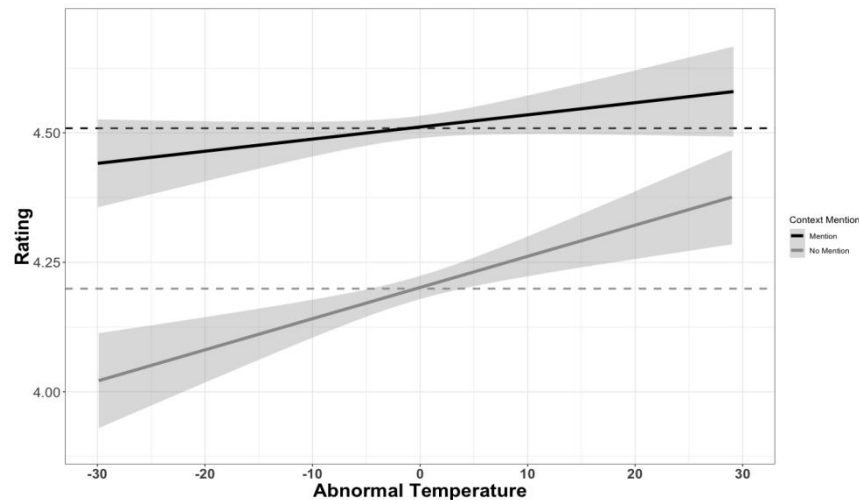
Mention of Context in Reviews: Analysis & Results

We hypothesize that context affects ratings because many reviewers misattribute aspects of their experience caused by context to the item being evaluated. A commonly identified cause of consumers’ misattributions is insufficient awareness (e.g., Kim, Park and Schwarz 2009; Schwarz and Clore 1983); thus, it seems plausible that context affects ratings if reviewers are not sufficiently aware of their context (specifically, its abnormality) or of the influence of context on experience. If so, we should see that reviews which mention context—indicating awareness—show a diminished effect of that context on ratings.

Author Accepted Manuscript

To test this, we measured whether a review mentioned experience-relevant context using GPT-4o-mini. We submitted each review for cold-weather gear to the model, with the prompt “*I will give you a product review. Provide one response. Does the review mention the weather in which the product was used? (1 if so, 0 if not). Respond with only a single integer (0-1)*”. GPT coded 41.9% of cold-weather gear reviews as mentioning context.⁸ We did not code reviews for non-cold-weather gear because weather is not directly experience-relevant for these products. Therefore, we investigated the simple effect of temperature on ratings for cold-weather gear.

Figure 2: Product Ratings for Cold-Weather Gear When Reviews Mention Context v. Not



Note: For this plot, we winsorize abnormal temperatures to be between -30 and 30 . This cuts the lowest 0.14% and highest 0.11% .

Results support the prediction that awareness decreases the effect of context on ratings. We estimated a version of our main model, predicting ratings with temperature, an indicator for whether a review mentioned context, and their interaction, with fixed effects for location, month, and product, and standard errors clustered by product. This found a negative interaction ($\beta = -$

⁸ As further checks, we asked GPT to code (for each review it coded as mentioning context) whether it discussed how the product performed in the weather, and how specifically the review mentioned context (0–9). 95.5% discussed how the product performed in the weather, supporting the notion that this mention is product specific. Context was rarely mentioned very specifically ($M = 4.06$), with 48.6% scoring between zero and three, and only 24.7% of reviews scoring above five.

.004, $t(14,048) = -3.348, p < .001$, 95% CI = $[-.006, -.001]$, Std. $\beta = -.051$; Figure 2), such that reviews not mentioning context showed a large effect of context ($\beta = .005, t(14,048) = 4.003, p < .001$, 95% CI = $[.003, .008]$, Std. $\beta = .076$), which is attenuated for reviews mentioning context ($\beta = .002, t(14,048) = 1.369, p = .171$, 95% CI = $[> -.001, .004]$, Std. $\beta = .025$). Thus, as with related studies of misattribution, we find attenuation with awareness of context. We also note that we observed a main effect of mentioning context on ratings, such that reviews that mention context at all are significantly more positive ($\beta = .373, t(14,050) = 19.285, p < .001$, 95% CI = $[.335, .411]$, Std. $\beta = .324$). We do not have a clear theoretical explanation for this, and thus avoid speculating on why it emerged.⁹

Discussion

Results from these data support our misattribution hypothesis (H_1) and provide preliminary evidence for our attenuation hypothesis (H_2). Ratings for products depend in part on the context in which the product is consumed. By comparing the effect of temperature on ratings between products whose purpose is to keep consumers warm and all other products, we are able to distinguish this experience-relevant context from more general misattribution of mood. We believe this both supports the notion that ratings measure consumer experiences, and also highlights a problem with that fact: Experience depends on context, and context is often extrinsic to the product being rated. Below, we discuss and clarify several practical implications of these findings, with specific opportunities for platforms introduced in the general discussion.

Bias from systematic differences in context between products

⁹ We conducted an exploratory analysis to assess possible mechanisms. We found reviews that mention context are significantly less likely to be one-stars (Std. $b = -.158$), and removing one-star ratings shrinks the interaction effect size by a considerable amount (Std. $b = -.003$ v. Std. $b = -.051$ with one-stars). This could indicate that reviewers who do not mention context are thinking less deeply about their experiences and rating more extremely.

Author Accepted Manuscript

It is likely that the effect of context on ratings harms some products on average, while helping others. Products that are systematically worn in colder than normal temperatures *should* receive lower (than deserved) ratings more frequently than products worn in warmer temperatures. However, we are able to glean only suggestive—not causal—evidence of this in our data, as the imperfect measures of quality we do have are somewhat confounded. For an illustrative example, REI.com includes a “Warmth rating” of either “Warm” ($N = 132$), “Warmer” ($N = 283$), or “Warmest” ($N = 108$) for 523 jackets in our data. Jackets rated “Warmest” are worn in colder temperatures ($M = 41.6$) than those rated “Warmer” ($M = 47.0$) and “Warm” ($M = 50.9$; $\beta = -6.587$, $t(8,146) = 16.441$, $p < .001$, 95% CI = $[-7.373, -5.802]$, Std. $\beta = -.377$), but there is no difference in the effect of temperature on ratings between “Warmest” jackets and others ($\beta < .001$, $t(7,340) = .424$, $p = .641$, 95% CI = $[-.003, .004]$, Std. $\beta = .011$). If this were causal, the effect of context on ratings would likely harm “Warmest” jackets due to consumers using them in unfavorable contexts. However, we caution over-interpretation of this result, which we present merely for illustrative purposes here.

Downstream effects of noise on rankings and search

This introduces noise into product rankings when sorted by rating. As an illustration, we ranked the top 30 men’s and women’s jackets in our data according to their average rating.¹⁰ Then, we used Model 5 to simulate a counterfactual for each individual rating as if it had

¹⁰ We broke ties in rankings by number of ratings, as is done on [REI.com](https://www.rei.com). We focus on 30 and 90 products because those are the default and maximum (respectively) number of products that can be viewed on a page at the time of writing, and most consumers do not search past the first page (Farronato, Fradkin, and MacKay 2023). If we consider all products in a category ($n = 305$ for men and $n = 293$ for women), the correlations are much larger ($r = .983$ for men and $r = .975$ for women). This apparent inconsistency stems from the fact that the change in rankings after the adjustment are localized (e.g., the 16th ranked product becomes the 29th ranked product), and that average ratings are much closer together among top-ranked products. The median magnitude of change across all products is 3 in both categories. The products at the very top and very bottom of the rankings change their relative positions, but they still stay near the top and bottom (respectively). This leads to a strong correlation across all products, but significant re-arrangement—and thus a low correlation—in any local window.

occurred in normal temperatures. To illustrate this process, we take an actual review from the data, in which a reviewer rated the Icebreaker Women's Descender winter jacket five stars in Finland, Minnesota on February 24, 2021. The three-day running average at this time was 34.870 degrees, which is 14.088 degrees colder than the average temperature on February 24 in Finland in other years in our data. Therefore, Model 5 predicts that the rating would have been .007 (14.088×0.0005) stars higher if the jacket had been worn in exactly average temperatures. We calculated adjusted ratings according to this process for each observation.

We then re-ranked the initial top 30 according to these adjusted ratings. The correlations between rankings on raw and adjusted ratings were quite low: $r = .096$ for men, $-.015$ for women. This was consistent among the top 90 ($r = .380, .186$), supporting our conjecture that noise in ratings can create substantial noise in rankings. These differences in rankings can affect how people search and what they ultimately choose. For example, Ursu (2018) identified a difference in click-through-rate between positions one and five on a search result page of over 100%, and identified gains in consumer welfare of nearly 25% from reduced noise.

Influence on ratings distributions

We also find that the effect of context on ratings increases noise in ratings' distributions. Specifically, more variation in the context in which a product is consumed is associated with a significant increase in the standard deviation of that product's ratings. To demonstrate, we summarized temperatures and ratings for each cold-weather product. We then regressed the standard deviation in ratings on the standard deviation in abnormal temperature, controlling for the number of ratings, and found a positive effect ($\beta = .012, t(1,372) = 2.435, p = .015, 95\% \text{ CI} = [.002, .022], \text{Std. } \beta = .066$). This suggests that products experienced in a wider range of contexts have more dispersion in their ratings distributions. Because this has been shown to reduce choice

Author Accepted Manuscript

by increasing consumers' uncertainty about product quality (He and Bond 2015; Matz and Wood 2005; Urbany, Dickinson, and Wilkie 1989), this may lead products consumed in a wider range of contexts to be particularly harmed by that context's effect on ratings.

Opportunities for platforms

These results also demonstrate opportunities for platforms. While we discuss others in the general discussion, the observation that the effect of context is attenuated when reviews mention context suggests that platforms may benefit by suggesting reviewers include context information in their reviews. At best, this could remove the effect of context on ratings. At worst, it could provide prospective consumers with useful information. However, we cannot make a causal claim about the effect of awareness on ratings in this study alone. Unlike variation in abnormal temperatures, whether a reviewer mentions context is not random, and attenuation could be due to differences between consumers. We address this in studies 3 and 4. Before doing so, we investigate a second instantiation of the effect of experience-relevant context in the same data.

Study 2: REI.com Ratings of Rain Jackets

In Study 2, we test generalizability by investigating another product category and form of experience-relevant context in our REI data—rain jackets and precipitation.

Data

These data come from the same set as Study 1. Here we investigate the effect of recent abnormal precipitation on products worn to keep consumers dry. Precipitation is measured in inches over a 24-hour period at the same stations as Study 1 temperature observations. Not every station reports precipitation every day. Of our total sample of 218,913 reviews, 58,088 (26.5%) have missing precipitation data. A further 1,443 (.7%) have precipitation reports with errors flagged by NCEI, leaving 159,382 reviews.

We collected 2,268 ratings for 119 men’s and women’s rain jackets. After removing 583 ratings with missing or flagged precipitation observations, we retain 1,685 ratings for 114 products. We compare these to the 157,697 ratings for other products that have precipitation observations. In sum, 52.6% of reviews were written on days that had zero precipitation recorded on that day and the two prior (42.1% for rain jackets, 52.7% for other), while the mean was .082 inches ($SD = .179$; $M_{Rain} = .109$, $SD_{Rain} = .207$; $M_{Other} = .082$, $SD_{Other} = .178$).

Main Analysis & Results

We test this prediction in a regression model similar to Model 5 in Table 1. Individual rating serves at the unit of analysis. We also include two focal variables and their interaction: Precipitation (in inches) over the last three days at the rater’s location, and an indicator variable for whether the rating is for a rain jacket (1 = rain jacket, 0 = not). The interaction coefficient reflects the differential effect of precipitation on ratings for rain jackets, and thus is the test of H_1 : A negative interaction coefficient suggests that precipitation affects ratings for rain jackets more than for other products—such that more precipitation leads to lower ratings. We also include fixed effects for location, month, and product, and cluster standard errors by product.

Results support our prediction that experience-relevant consumption impacts ratings. A negative interaction coefficient indicates that ratings for rain jackets vary depending on consumption context ($\beta = -.407$, $t(141,691) = -2.334$, $p = .020$, 95% CI = $[-.749, -.065]$, Std. $\beta = -.058$). There is no simple effect of precipitation for other products ($\beta = -.019$, $t(141,691) = -1.118$, $p = .264$, 95% CI = $[-.054, .015]$, Std. $\beta = -.003$). Re-coding this model such that the indicator equals zero for rain jackets reveals that with more precipitation, ratings are lower for

Author Accepted Manuscript

products whose purpose is to keep consumers dry ($\beta = -.426$, $t(141,691) = -2.458$, $p = .014$, 95% CI = $[-.766, -.086]$, Std. $\beta = -.061$).

Robustness

We again assessed robustness through a specification curve analysis, varying each combination of all potential analytic choices, yielding 6,912 models. These analytic choices are the same as in Study 1 with two exceptions. First, there are only three context measures (precipitation over a one, three, or seven-day running average), as precipitation is only measured in daily totals. Second, rather than varying the inclusion of a control for precipitation, we vary the inclusion of a control for temperature. Results (web appendix C) demonstrate negative interaction estimates in 6,890 (99.7%) of models, with 95% confidence intervals not overlapping with zero in 1,792 (25.9%). This smaller proportion may be due to the smaller sample of reviews for rain jackets, as well as the fact that precipitation has relatively little variation.

Text Analysis

We also repeated similar text analyses as Study 1. The proportion of reviews that mention context for rain jackets is higher (60.3%) in Study 2 than it was for cold-weather gear in Study 1 (41.9%). However, we found no interaction between product type and precipitation on sentiment ($\beta = -.692$, $t(81,142) = -1.526$, $p = .127$, 95% CI = $[-1.581, .197]$, Std. $\beta = -.032$). We also found no attenuation of the effect among reviews that mentioned context ($\beta = .103$, $t(1,225) = .208$, $p = .836$, 95% CI = $[-.879, 1.085]$, Std. $\beta = .015$). This is likely due to a combination of the smaller sample, as well as the aforementioned issues with precipitation.

Discussion

Investigates the effect of experience-relevant context in a second scenario in the same setting. In fact, the standardized effect size is slightly larger in magnitude for precipitation than temperature in Study 1 (Std. $\beta = -.058$ v. $.042$). However, this analysis is secondary because the sample of ratings for rain jackets is small, and precipitation is not normally distributed. This reduces our ability to observe effects on review sentiment, or attenuation by mention of context.

Overview of Experimental Studies

Studies 1 and 2 establish evidence for the effect of experience-relevant context on real ratings. In the remaining studies, we adopt an experimental paradigm to study our hypotheses. This allows us to randomly assign participants to treatments and thus more clearly establish a causal relationship between context and ratings. Further, it provides the opportunity to (i) assess generalization by examining a greater array of products (energy drinks, hotel beds, restaurants, running shoes, and sunscreen) and contexts (fatigue, hunger, running distance, and UV index), (ii) further probe the role of contextual awareness in the misattribution process, and (iii) test potential interventions platforms can use to improve their collection of ratings.

Study 3: Extension Across Product Categories

Study 3 extends the previous findings in two key ways. First, we test generalizability by examining five new product categories (and contexts) in addition to winter jackets (and temperature): energy drinks (alertness), hotel beds (need to sleep), restaurants (hunger), running shoes (distance ran), and sunscreen (UV index).¹¹ Participants read about their experience with one of these products, learn about the context in which they had this experience (or not in the “absent” condition), and then provide a rating for the product.

¹¹ Web Appendix Study 1 (Web Appendix D) replicates the winter jackets replicate from this study.

Author Accepted Manuscript

Second, we experimentally test our process argument—that context influences ratings because consumers misattribute aspects of their experience to products that are caused by context. We do so by manipulating the degree to which we make the context and its likely effect on experience salient to the participant.

Participants & Procedure

We recruited 903 CloudResearch-approved participants to participate in this study study. Of these, 10 started the experiment more than once. Consistent with our preregistration, we removed their responses, leaving us with 893. Of these, 13 did not complete all trials. Because we did not preregister to exclude them, we keep their responses in the analysis presented here.



Participants completed six trials. In each, they were presented with an imagined product experience (e.g., *“Imagine you recently purchased a new winter jacket, going to work, the grocery store, running errands, and walking outside”*). Within participants, we varied whether the experience was either as expected or worse than expected in randomized order, such that participants always saw three replicates from each of the two conditions.

Between-subjects, participants were randomly assigned to one of three “context awareness” conditions: “absent,” “light,” or “heavy.” These were designed to increase awareness of context and thus provide a test of mechanism. The more aware participants are of context, the less it should affect their ratings. We note that the influence of context in this study was always consistent with their experience: If their experience was as expected, the contextual influence was favorable (e.g., warm weather for a jacket). If their experience was worse than expected, the contextual influence was unfavorable (e.g., cold weather for a jacket).

In the absent condition, participants only read about their experience and received no information about the contextual factors that might have influenced their experience. This

condition should serve as a baseline for comparison, as participants have no ability to correct for the influence of context on experience, because they have no information about that context. In the light awareness condition, participants saw information about the context that led to their rating (e.g., for winter jackets in the “as expected” condition: “While you wore this winter jacket, the average temperature was slightly warmer than you would normally expect it to be at this time of year”). In this condition, participants have the ability to correct for the contextual influence, but are not explicitly reminded they should do so. Lastly, in the heavy awareness condition participants saw the same information as in the other conditions, but also a message emphasizing the potential the influence of this context (e.g., “This was probably the reason that you felt as warm (you did not feel as warm) as you had hoped when you bought the winter jacket.”). Further details are provided in Figure 3 (sample stimuli) and Table 5 (message details).

Figure 3: Example Stimuli for Winter Jackets Replicate

Worse Condition	As Expected Condition
<p>Product: Winter Jacket</p> 	<p>Product: Winter Jacket</p> 
<p>Your experience</p> <p>After your new winter jacket arrived, you wore the product consistently for a few days. You used this winter jacket as you normally would, going to work, the grocery store, running errands, and walking outside.</p> <p>While you wore this winter jacket, you noticed you did not feel as warm as you had hoped when you bought the winter jacket.</p> <p>While you wore this winter jacket, the average temperature was slightly colder than you would normally expect it to be at this time of year. This was probably the reason that you did not feel as warm as you had hoped when you bought the winter jacket.</p> <p>You liked the look of this winter jacket.</p>	<p>Your experience</p> <p>After your new winter jacket arrived, you wore the product consistently for a few days. You used this winter jacket as you normally would, going to work, the grocery store, running errands, and walking outside.</p> <p>While you wore this winter jacket, you noticed you felt as warm as you had hoped when you bought the winter jacket.</p> <p>While you wore this winter jacket, the average temperature was slightly warmer than you would normally expect it to be at this time of year. This was probably the reason that you felt as warm as you had hoped when you bought the winter jacket.</p> <p>You liked the look of this winter jacket.</p>

Note: Highlights were not presented to participants. Red highlights indicate information added in the “light” and “heavy” awareness conditions. Blue highlights were only added in the “heavy” condition.

After reading about their experience and the context (or not, in the absent condition), participants rated the product on a 1–10 (poor–excellent) scale. In the light awareness condition, participants were shown the same context information from the page before. In the heavy awareness condition, we added “You should not let this {context} impact your rating”

Author Accepted Manuscript

immediately before the rating to further reinforce awareness. Participants completed six product trials in random order. This creates a 2 (experience [within]: as expected, worse) \times 3 (context awareness [between]: absent, light, heavy) \times 6 (product category [within]) mixed design.

Because some participants did not complete all trials, the total number of ratings was 5,306.

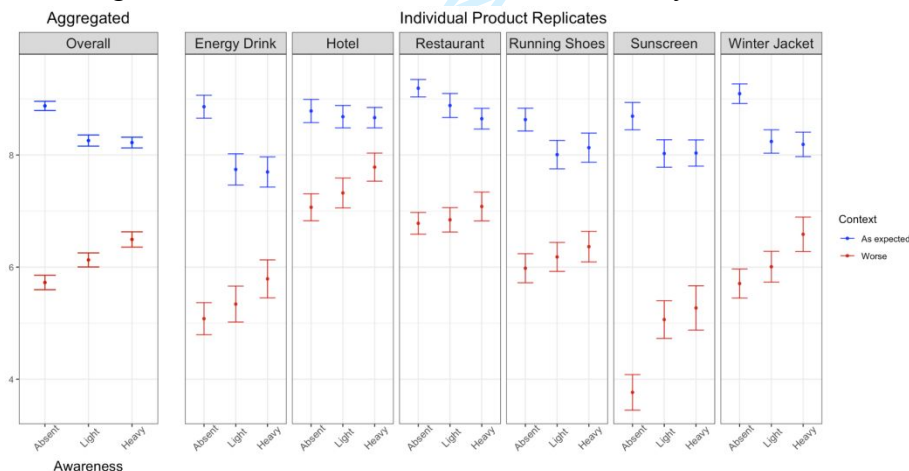
Table 5: Product and Context Replicates in Study 3.

Product	Worse Experience	As Expected Experience	Context
Energy drink	Not as alert, nor energized as expected	As alert and energized as expected	Tiredness before drink (very vs not at all)
Hotel	Slightly worse sleep than expected	Slightly better sleep than expected	Tiredness before bed (not at all vs very)
Restaurant	Liked, but did not love the food	You loved the food	Hunger before eating (not at all vs very)
Running shoes	Did not feel quite as much cushion as you had hoped	Felt about as much cushion as you had hoped	Running distance (more vs less) than usual
Sunscreen	Noticed a sunburn begin to appear	No obvious sunburn	UV index (higher vs lower)
Winter jacket	Did not feel as warm as you had hoped	Felt as warm as hoped	Average temperature (higher vs lower)

Note: For all products, the heavy reminder contained this information, as well as "This is why your experience...". This information was also present on the rating page.

Analysis & Results

Figure 4: Mean Ratings and 95% Confidence Intervals from Study 3 Across Product Replicates.



Note: Error bars represent 95% confidence intervals.

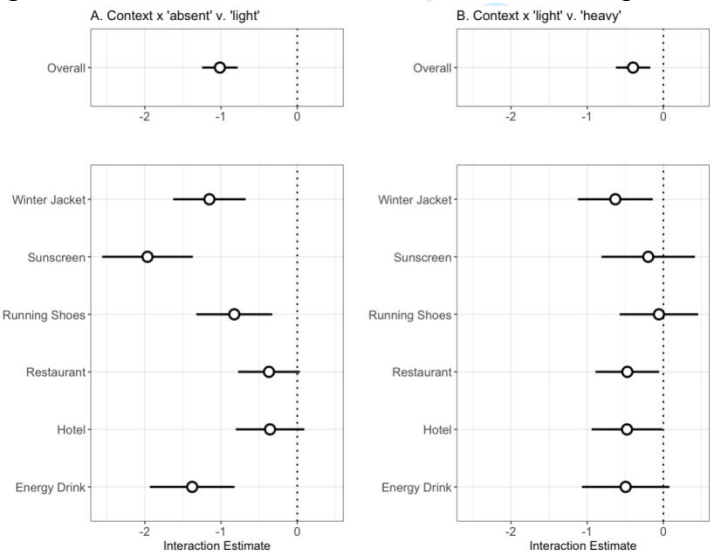
Our focus is on the attenuating effect of increasing awareness: The difference in ratings between “as expected” and “worse” experience should decrease as awareness of context increases. As shown in Figure 4, this is the pattern we observe across the different product categories: The rating difference is the largest when no context information is provided (absent

condition) and is the smallest when we provide the most explicit information about the influence of context (heavy condition).

We analyze this data using a linear regression. We run a model predicting ratings with the following predictor variables: dummy codes for the light and heavy awareness conditions, a contrast code for context, and interactions of context and awareness conditions. We included participant fixed effects and clustered standard errors by participant.

The simple effect of context in the absent awareness condition revealed higher average ratings after “as expected” experiences than after worse experiences ($\beta_{Context} = 3.151, t(4,415) = 33.577, p < .001, 95\% \text{ CI} = [2.967, 3.335], \text{Std. } \beta = 1.499$). However, this effect was significantly attenuated by awareness of context. We found an interaction of context and the light awareness condition ($\beta_{Context \times Light} = -1.019, t(4,415) = -7.997, p < .001, 95\% \text{ CI} = [-1.269, -.769], \text{Std. } \beta = -.485$; Figure 5A) and of context and the heavy condition ($\beta_{Context \times Heavy} = -1.424, t(4,415) = -11.162, p < .001, 95\% \text{ CI} = [-1.674, -1.173], \text{Std. } \beta = -.677$).¹²

Figure 5: Interaction Estimates Across Product Replicates



Note: Lines represent 95% confidence intervals.

¹² These results are not changed if we control for product (Std. $\beta_{Context \times Light} = -.483$; Std. $\beta_{Context \times Heavy} = -.672$).

Author Accepted Manuscript

Re-coding the model to compare the light and heavy awareness conditions found that the additional explanation of the impact of context on experience created significantly further attenuation beyond mere awareness we found a significant interaction of context and the heavy (v. light) awareness condition ($\beta_{Context \times Heavy} = -.404$, $t(4,415) = -3.313$, $p < .001$, 95% CI = $[-.644, -.165]$, Std. $\beta = -.192$; Figure 5B). These interaction results were consistent across product replicates, as shown in Figure 5—which illustrates interaction estimates within each product replicate. Despite being powered to test the overall effect (rather than the effect within each replicate), we find attenuation between the absent and light conditions in four of six products, and between the light and heavy conditions in three.

Discussion

Study 3 provides experimental evidence that ratings for a wide variety of product categories may be influenced by consumption context. It also provides evidence that this influence is likely unintentional and reflects a failure to consider the role of context on consumption experience: When we increase the salience of the consumption context, participants provide ratings that are less influenced by it. In Study 4 we build on this insight—that reminding raters to consider context can reduce the influence of that context on their ratings—to develop and assess practical interventions platforms can potentially employ when soliciting ratings.

Study 4: Debiasing Through Prompts

In Study 4, we assessed the effectiveness of four different prompts that platforms can use when soliciting ratings. The goal of these prompts was to reduce the degree to which context influenced ratings, by increasing awareness of the context at the time of rating. We explore the relative effectiveness of prompts that are more general (e.g., only reminding of the context) and more specific (e.g., referring to the specific context that affected their experience), with the idea

Author Accepted Manuscript

that general prompts could be implemented more broadly (i.e., are neither product nor context specific) but specific prompts might be more effective.

Participants and Procedure

We recruited 1,405 participants from Connect to complete this study. Six failed an attention check and, following our preregistration, were removed from the data, leaving 1,399. Study 4 employed a fully between-subject design and only used a single product replicate: winter jackets. We note that unlike Study 3, we provided context information to all participants.

Participants were randomly assigned to one of two experience conditions—as expected or worse. In the as expected condition, participants read *“After your new jacket arrived, you wore the product consistently for a few days. You used this jacket as you normally would. While you wore this jacket, you felt warm enough. Compared to the temperature you would normally wear this jacket in, these few days were warm. You liked the look of this jacket.”* In the worse condition, participants read *“After your new jacket arrived, you wore the product consistently for a few days. You used this jacket as you normally would. While you wore this jacket, you felt slightly cold. Compared to the temperature you would normally wear this jacket in, these few days were cold. You liked the look of this jacket.”*

After learning about their experience (and the context that influenced it), participants were told that the store they had purchased their jacket from asked them to evaluate the purchase. Between-subjects, we randomly assigned participants to one of five rating prompt conditions. The first (baseline) condition most closely replicates standard rating prompts by providing and prompting no context information. In this condition, participants only answered the question *“Thinking about the jacket as objectively as possible, how would you rate this jacket?”* (1–10; poor–excellent). Participants in the other four conditions completed the study by responding to

Author Accepted Manuscript

the same question, after being exposed to a condition-specific debiasing prompt. The second (context-general) condition asked participants two free-response questions before eliciting their rating (*“In what context did you wear this jacket? Was this context abnormal in any way?”* and *“Do you think this context impacted your feelings about the jacket in any way? If so, how?”*). These questions were designed to elicit consideration of the effect of context in as free of a format as possible. The third (context-weather) condition mirrored the second, but replaced “context” with “weather,” to assess whether specifying the relevant context is necessary.

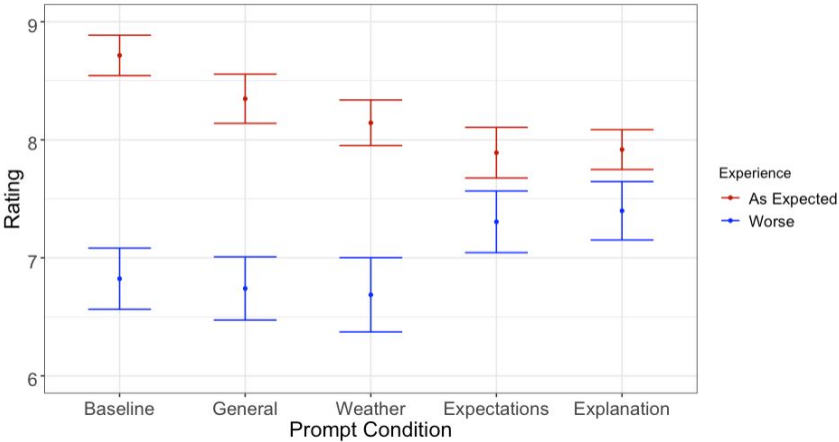
The final two conditions tested more direct prompts. The fourth (context-expectation) began the same as the third condition (*“In what weather did you wear this jacket? Was this weather abnormal in any way?”*), but then directly explained the effect of context (*“Many people under-estimate the influence of things like weather on their experiences. For example, colder temperatures will always make you feel cold, no matter the quality of the jacket. Warmer temperatures will always make you feel warm for the same reason.”*). Finally, the fifth (context-explanation) condition did not ask about weather, but directly reminded participants *“Heads up! When you wore this jacket, it was especially {warm/cold} where you are”* before explaining the effect of this context. These final two conditions are distinguished from the others by increasing awareness not only in the context itself, but in the effect of context on experience.

Analysis & Results

As shown in Figure 6, we observe the largest difference in ratings with the baseline prompt. The other prompts attenuated this difference, suggesting that platforms could employ these to achieve less context-influenced ratings. To analyze these results, we first evaluated whether differing prompts attenuated the effect *in general* by testing the omnibus prompt \times experience interaction in ANOVA, with rating as the dependent variable. This revealed a large

omnibus interaction effect ($F(4, 1389) = 13.893, p < .001$), suggesting that the effect of context differed after different rating prompts.

Figure 6: Mean Ratings and 95% Confidence Intervals from Study 4 Across Prompts.



Note: Error bars represent 95% confidence intervals.

Next, we investigated specific comparisons of each prompt condition to the baseline. To do so, we constructed a series of dummy codes for each prompt condition. We then analyzed pairwise comparisons of baseline and prompt condition in four separate linear regressions with ratings as the outcome, and the prompt condition, experience condition ($-.5$: worse, $.5$: as expected), and their interaction as predictors.

These regressions revealed a large effect of context in the baseline condition ($\beta = 1.893, t(556) = 11.585, p < .001, 95\% \text{ CI} = [1.572, 2.213], \text{Std. } \beta = 1.146$). This was not reduced by either the context-general prompt ($\beta_{\text{Interaction}} = -.285, t(556) = -1.201, p = .230, 95\% \text{ CI} = [-.751, .181], 95\% \text{ CI} = [-.644, -.165], \text{Std. } \beta = -.173$) or the context-weather prompt ($\beta_{\text{Interaction}} = -.435, t(557) = -1.748, p = .081, 95\% \text{ CI} = [-.924, .054], \text{Std. } \beta = -.256$). However, we found attenuation by adding the expectation prompt ($\beta_{\text{Interaction}} = -1.307, t(569) = -5.492, p < .001, 95\% \text{ CI} = [-1.774, -.840], \text{Std. } \beta = -.824$) and explanation prompt ($\beta_{\text{Interaction}} = -1.374, t(586) = -6.189, p < .001, 95\% \text{ CI} = [-1.810, -.938], \text{Std. } \beta = -.910$).

Author Accepted Manuscript

These debiasing prompts appear to have had similar effects in both experience conditions. Treating the debiasing prompt conditions as a continuous linear predictor, we estimated the simple effect of prompt in each condition. This revealed similar effects (of opposite sign) in the as expected ($\beta = -.202$, $t(1,395) = -5.427$, $p < .001$, 95% CI = $[-.245, -.129]$, Std. $\beta = -.186$) and worse ($\beta = .167$, $t(1,395) = 4.481$, $p < .001$, 95% CI = $[.094, .240]$, Std. $\beta = .154$) conditions.

Discussion

This experiment demonstrates potential strategies for platforms to employ when eliciting ratings from consumers. We find that prompts which make consumers aware of the specific influence of their context on their experience—here by explaining that influence—can substantially reduce the effect of context on ratings. Additionally, this experiment demonstrates the insufficiency of merely highlighting consumers' context in a general sense (or prompting them to consider it generally), which led to much smaller decreases in the effect of context. This suggests that the attenuation in Study 1 among reviews that mention context was not merely due to awareness of context, but also of awareness of the *influence* of context on experience.

It is noteworthy that we observe similar attenuation effect sizes from both experience conditions. We believe this is consistent with our framework, as both conditions provided participants with a directional (positive or negative) experience, and equivalent prompts to explain this experience. Highlighting the salience of context appears to encourage participants to attribute some of both their positive and negative experiences to context (v. the product).

General Discussion

This research investigates the influence of experience-relevant consumption context on user-generated ratings, finding that ratings are impacted by context in ways users of ratings

cannot observe. We find evidence for this in analysis of real ratings from two product categories and contexts in the field, in which we are also able to distinguish effects of experience from mood. We also find attenuation in reviews for cold-weather products that mention context, suggesting that awareness of context might decrease its effect on ratings. This evidence is replicated and extended in two experiments, where we also demonstrate strategies for platforms.

This research contributes to the literatures on user-generated ratings and the role of context in consumer judgment and decision-making. Our focus on experience-relevant context influencing consumption directly can be distinguished from past research. Most related research studies explicitly how context impacts evaluations by affecting mood (e.g., Brandes and Dover 2022; Cohen et al. 2018; Schwarz and Clore 1983). Other research considers a different role of context, and does not distinguish context’s impact on experience from mood. For instance, Figini, Leoni and Vici (2024) studied the role of surprise on ratings, with the difference between actual and forecasted weather as a measure of surprise. They found that sunnier than forecasted weather increased ratings for hotels, and vice versa for surprisingly rainy weather. While Figini et al. (2024) focused on surprise and mis-forecasting as influences on ratings, their result is consistent with our prediction. However, they only considered a single consumption experience (hotels), and could not discern an effect of context on consumption experience from mood.¹³

Research in economics has also demonstrated systematic effects of environmental variation on other consumer behaviors (e.g., Busse et al. 2015; Conlin, O’Donoghue, and Vogelsang 2007; Haggag et al. 2019). A literature on projection bias indicates that people incorrectly project their current context will persist, leading them to make purchase decisions that are overly tailored to present circumstances. For example, Busse et al. (2015) found that

¹³ We also consider a range of durable goods, and do not require consumers to be aware of forecasts.

Author Accepted Manuscript

consumers were more likely to purchase convertible cars on sunny days. They propose this is because driving a convertible is more enjoyable on a sunny day, and that consumers neglect the uniqueness of sunny days in the future—a similar, but distinct, lack of awareness of context.

Practical Implications

The present research has a range of important implications for platforms, manufacturers, and consumers. While the effect size we quantify in our analyses of REI.com ratings may appear small in comparison to those derived from experimental paradigms (Std. $\beta = .032$ in Study 1, $.058$ in Study 2), it is important to contextualize this measure. Our inexact measure of consumption context introduces random error to our model. This diminishes the measured effect size by definition, although we cannot know by how much.

Increased variance in ratings

Beyond any effect on average ratings, increased variation in ratings will make consumers more negative about products for at least two reasons. First, consumers attend to negative information when learning about products and reading ratings (Chevalier and Mayzlin 2006; Gottschalk and Mafael 2017), which is consistent with a broader human bias for negative information (Rozin and Royzman 2001). Any increase in negative information will be more impactful than an equivalent increase in positive information, as consumers will attend to the negative information more than the positive. Therefore, there does not need to be a large effect on average ratings to have a large effect on average consumer perceptions—an increase in noise alone will lead to an increase in the number of negative reviews, which are more impactful.

The second reason that the effect we observe may make consumers more negative is because higher variance in ratings increases consumer uncertainty. Prior research shows that variation in ratings generally makes consumers less positive about products (Meyer 1981; Yin,

Mitra, and Zhang 2016) less confident in reviews’ helpfulness (Lee, Lee, and Baek 2021), and less likely to make a choice (Lee et al. 2021; Varga and Albuquerque 2019; Zhu and Zhang 2010). This negative effect is stronger when consumers see variation to come from variation in quality (He and Bond 2015), which is the case in the scenarios we study (e.g., felt warmth, dryness). Thus, the influence of experience-relevant context is likely to be detrimental to choice.

Variance in ratings also brings noise to product rankings, and consumer search as a result. This relies on the fact that most products are not rated frequently enough for noise caused by random variation from a few raters’ contexts to be canceled out. The median product on REI.com has six ratings, and 57.7% have fewer than ten. This is not unique to our data, as other work has identified small samples as a significant barrier to ratings’ validity on other sites (de Langhe et al. 2016), while consumers treat average ratings as a more diagnostic cue of quality than sample size (Watson et al. 2018). Therefore, each additional rating has a large impact on the ordering of products when ranked by average rating. This is illustrated in the discussion to Study 1, where we calculated an “adjusted” rating for each rating of each jacket in our data, and compared product rankings from these adjusted ratings to rankings from the raw ratings. This highlights the outsized effect that seemingly small differences in ratings can have on downstream consequences, such as consumer search. This is especially true as Ursu (2018) demonstrated large impacts on consumer search and welfare from small differences in ranking—click through rates for the top listed product in her data are over twice those of the fifth listed product.

Implications for consumers

While detractors have argued ratings are invalid because they don’t reflect objective quality (e.g., de Langhe et al. 2016), advocates have suggested that this is a chief benefit, as objective quality does not capture the entire consumption experience (Simonson 2016).

Author Accepted Manuscript

However, there is no requirement for raters to only consider things that will be relevant to others, and experience includes irrelevant factors (e.g., idiosyncratic temperature). Moreover, reviews that do not mention context are most influenced by context. If the opposite were true, the effect of experience-relevant context on ratings could be seen as a boon for consumers—if affected reviews also mentioned context, consumers could learn how a product performs in different circumstances, leading them to better meet their needs.

Potential Solutions

A clear solution is identified in Study 4. Platforms should identify relevant context effects and prompt ratings with messages specifically addressing those effects. This should attenuate the effect of context on ratings. While we found only small attenuation in Study 4 among the two conditions that did not explain the effect of context, doing so is likely better than the status quo, as these prompts would increase the number of reviews mentioning context.

The observation in Study 1 that the effect of context is attenuated among reviews that mention context also suggests changes for review solicitation. Platforms could provide additional incentives and rewards to consumers who provide more detailed reviews—such as those who mention weather when reviewing cold-weather products. Doing so would capitalize on the fact that at least some of the attenuation in Study 1 is likely due to differences between reviewers. In this view, our findings suggest a way for platforms to distinguish reviewer quality.

This is not the limit of ways platforms could use context information either. Platforms could perform our same analyses, including the adjustment of ratings we describe above. Platforms could then provide the context information that many reviewers leave out (e.g., listing the temperature a user consumed their product in), or could even directly adjust ratings by quantifying and removing the effect of contexts. Some retailers (e.g., Amazon.com) already

perform corrections on the aggregate ratings they display (Matsakis 2019), though these corrections are not specific to consumption context’s effects on ratings.

However, as directly altering ratings might elicit concern on the part of consumers (that platforms are “tipping the scales” in some way),¹⁴ we suggest using context information to inform recommendation systems and ranking algorithms. Specifically, using context information from existing reviews could help platforms to identify what products perform best in different contexts, leading to tailored recommendations for specific consumers. In this way, the solution is not to remove the effect of context, but to use it as information for other decision aids for consumers. Prior research on rankings and consumer search supports the positive benefits of doing so. For example, Ursu (2018) identifies consumer welfare gains of nearly 25% from less noisy ranking algorithms, while utility-based rankings (which context information could inform) lead to higher platform revenues (De los Santos and Koulayev 2017; Ghose et al. 2014), and consumer welfare is increased when consumers are informed about utility-based rankings (Chen and Yao 2017). Utilizing context’s influence on ratings—rather than attempting to remove it—is also a simpler challenge, as it does not require platforms to debias consumers’ judgments.

Future Research

We hope that the current research sparks interest in context effects in ratings. Future work could make use of more certainly measured contexts to more accurately quantify the effect size, which is likely underestimated here. Likewise, future research could expand upon the differences in mitigation by awareness we observed in Study 3 between products. It appears that the effects of context on potentially “simpler” products (sunscreen and energy drinks) were most strongly mitigated by the “light” awareness condition (i.e., informing participants of their

¹⁴ We thank an anonymous reviewer for raising this consideration.

Author Accepted Manuscript

context, without discussing its influence). This may be because the connection between context and experience is easiest to understand for simple products. This would suggest that the mitigation strategies tested in Study 4 would be most effective for such products.

We also hope to see future research engages with the use of context information in product recommendation systems. Platforms could use this information strategically. For example, platforms should be able to specifically recommend the best jacket for a specific consumer by analyzing the ratings of other consumers in similar climates. Platforms should also be able to use this information to segment reviewers according to the richness and objectivity of their reviews, rather than simply rewarding the frequency of reviewing.

Limitations and Constraints to Generality

Limitations relate primarily to our observational data collected from REI.com. These data do not allow us to have a measure of “objective quality” against which to compare the observed results. Other research has been able to investigate product categories with such measures (e.g., de Langhe et al. 2016), though without a clear measure of consumption context. Our products also limit scope to context created by weather; although we believe our results extend beyond such contexts, we cannot directly test that here. A limitation to generality comes through our experiments, which test six experiences in Study 3, and one in Study 4. Thus, mitigation strategies in Study 4 are context-specific, and generalization will require alterations.

Conclusion

This research adds to a growing literature on user-generated ratings, which suggests that user-generated ratings have tremendous potential but are clearly flawed. In this paper, we demonstrate one such flaw: User-generated ratings seem to reflect user experience, but experience is partly idiosyncratic.

References

Achiam, Josh, et al. (2023), “Gpt-4 technical report.” *arXiv preprint arXiv:2303.08774*.
Accessed March 25, 2025.

Abdurahman, Suhaib, Alireza Salkhordeh Ziabari, Alexander Moore, Daniel Bartels, and Morteza Dehghani (2024), “Evaluating large language models in psychological research: A guide for reviewers,” *Advances in Methods and Practices in Psychological Science*.

Bambauer-Sachse, Silke and Sabrina Mangold (2011), “Brand equity dilution through negative online word-of-mouth communication,” *Journal of Retailing and Consumer Services*, 18, 38–45.

Belal, Mohammad, James She, and Simon Wong (2023), “Leveraging chatgpt as text annotation tool for sentiment analysis.” *arXiv preprint arXiv:2306.17177*.

Brandes, Leif and Yaniv Dover (2022), “Offline Context Affects Online Reviews: The Effect of Post-Consumption Weather,” *Journal of Consumer Research*, 49 (4), 595–615.

Busse, Meghan R., Devin G. Pope, Jaren C. Pope, and Jorge Silva-Risso (2015), “The Psychological Effect of Weather on Car Purchases,” *The Quarterly Journal of Economics*, 130 (1), 371–414.

Chen, Pei-Yu, Yili Hong, and Ying Liu (2018), “The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments,” *Management Science* 64 (10), 4629–4647.

Chen, Yubo, Qi Wang, and Jinhong Xie (2011), “Online Social Interactions: A Natural Experiment on Word of Mouth versus Observational Learning,” *Journal of Marketing Research*, 48 (2), 238–54.

Author Accepted Manuscript

- Chen, Yuxin and Song Yao (2017), "Sequential search with refinement: Model and application with click-stream data," *Management Science*, 63 (12), 4345-4365.
- Chevalier, Judith A. and Dina Mayzlin (2006), "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 43 (3), 345-54.
- Chintagunta, Pradeep K., Shyam Gopinath, and Sriram Venkataraman (2010), "The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets," *Marketing Science*, 29 (5), 944-57.
- Cohen, Joel B., Michelle T. Pham, and Eduardo B. Andrade (2018), "The nature and role of affect in consumer behavior," In *Handbook of consumer psychology* (pp. 306-357). Routledge.
- Conlin, Michael, Ted O'Donoghue, and Timothy J. Vogelsang (2007), "Projection Bias in Catalog Orders," *The American Economic Review*, 97 (4), 1217-49.
- de Langhe, Bart, Philip M. Fernbach, and Donald R. Lichtenstein (2016), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*, 42 (6), 817-33.
- De los Santos, Babur and Sergei Koulayev (2017), "Optimizing click-through in online rankings with endogenous search refinement," *Marketing Science*, 36 (4), 542-564.
- Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken (2014), "What do we learn from the weather? The new climate-economy literature," *Journal of Economic Literature*, 52(3), 740-98.
- Dellarocas, Chrysanthos, Xiaoquan (Michael) Zhang, and Neveen F. Awad (2007), "Exploring the value of online product reviews in forecasting sales: The case of motion pictures," *Journal of Interactive Marketing*, 21 (4), 23-45.

Author Accepted Manuscript

- Dutton, Donald G., and Arthur P. Aron (1974), "Some Evidence for Heightened Sexual Attraction Under Conditions of High Anxiety," *Journal of Personality and Social Psychology*, 30(4), 510.
- Farronato, Chiara, Andrey Fradkin, and Alexander MacKay (2023), "Self-Preferencing at Amazon: Evidence from Search Rankings," *SSRN Working Paper*, <https://ssrn.com/abstract=4331880> or <http://dx.doi.org/10.2139/ssrn.4331880>
- Figini, Paolo, Veronica Leoni, and Laura Vici (2024), "And suddenly, the rain! When surprises shape experienced utility." *Journal of Economic Behavior & Organizations*, 224, 771-784.
- Ghose, Anindya, Panagiotis G. Ipeirotis, and Beibei Li (2014), "Examining the impact of ranking on consumer behavior and search engine revenue," *Mgmt Science*, 60(7), 1632-1654.
- Gilbert, Daniel T. and Patrick S. Malone (1995), "The correspondence bias," *Psychological Bulletin*, 117 (1), 21-38.
- Gottschalk, Sabrina A and Alexander Mafael (2017), "Cutting through the online review jungle: Investigating selective eWOM processing," *Journal of Interactive Marketing*, 37, 89-104.
- Haggag, Kareem, Devin G. Pope, Kinsey B. Bryant-Lees, and Maarten W. Bos (2019), "Attribution bias in consumer choice," *The Review of Economic Studies*, 86 (5), 2136-83.
- He, Stephen X. and Samuel D. Bond (2015), "Why Is the Crowd Divided? Attribution for Dispersion in Online Word of Mouth," *Journal of Consumer Research*, 41 (6), 1509-27.
- Imschloss, Monika and Christina Kuehnle (2019), "Feel the music! Exploring the cross-modal correspondence between music and haptic perceptions of softness," *Journal of Retailing*, 95(4), 158-169.

Author Accepted Manuscript

- Jones, Edward E., and Victor A. Harris (1967), "The Attribution of Attitudes," *Journal of Experimental Social Psychology*, 3(1), 1–24.
- Hakkyun Kim, Kiwan Park, and Norbert Schwarz (2010), "Will This Trip Really Be Exciting? The Role of Incidental Emotions in Product Evaluation," *Journal of Consumer Research*, 36 (6), 983–991, <https://doi.org/10.1086/644763>
- Lee, Soyeon, Saerom Lee, and Hyunmi Baek (2021), "Does the dispersion of online review ratings affect review helpfulness?" *Computers in Human Behavior*, 117, 106670.
- Matsakis, Louise (2019), "What Do Amazon's Star Ratings Really Mean?" Accessed July 2 2024. <https://www.wired.com/story/amazon-stars-ratings-calculated/>
- Matz, David C., and Wendy Wood (2005), "Cognitive dissonance in groups: the consequences of disagreement." *Journal of Personality and Social Psychology*, 88 (1), 22–37.
- Meister, Matt and Nicholas Reinholtz (2025), "Quality Certifications Influence User-Generated Ratings," *Journal of Consumer Research*, ucaf008. <https://doi.org/10.1093/jcr/ucaf008>
- Meyer, Robert J. (1981), "A Model of Multiattribute Judgments under Attribute Uncertainty and Informational Constraint," *Journal of Marketing Research*, 18 (4), 428–41.
- Meyers-Levy, Joan, Rui Zhu, and Lan Jiang (2010), "Context effects from bodily sensations: Examining bodily sensations induced by flooring and the moderating role of product viewing distance," *Journal of Consumer Research*, 37 (1), 1–14.
- Oliver, Richard L. (1977), "Effect of Expectation and Disconfirmation on Postexposure Product Evaluations: An Alternative Interpretation," *Journal of Applied Psychology*, 62 (4), 480–486.
- Park, Sungsik, Woochoel Shin, and Jinhong Xie (2021), "The Fateful First Consumer Review," *Marketing Science*, 40 (3), 481–507.

Author Accepted Manuscript

Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjeh, Claire E. Robertson, and Jay J.

Van Bavel (2024). "GPT is an effective tool for multilingual psychological text analysis."

Proceedings of the National Academy of Sciences, 121 (34), e2308950121.

Ross, Lee (1977), "The Intuitive Psychologist And His Shortcomings: Distortions in the

Attribution Process," *Advances in Experimental Social Psychology*, 10, 173–220.

Rozin, Paul and Edward B. Royzman (2001), "Negativity Bias, Negativity Dominance, and

Contagion," *Personality and Social Psychology Review*, 5 (4), 296–320.

Schwarz, Norbert and Gerald L. Clore (1983), "Mood, misattribution, and judgments of well-

being: Informative and directive functions of affective states," *Journal of Personality and*

Social Psychology, 45 (3), 513–523.

Schoenmueller, Verena, Oded Netzer, and Florian Stahl (2020), "The polarity of online reviews:

Prevalence, drivers and implications," *Journal of Marketing Research* 57 (5), 853-877.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson (2020), "Specification curve analysis,"

Nature Human Behaviour, 4 (11), 1208–14.

Simonson, Itamar (2016), "Imperfect Progress: An Objective Quality Assessment of the Role of

User Reviews in Consumer Decision Making, A Commentary on de Langhe, Fernbach,

and Lichtenstein," *Journal of Consumer Research*, 42, 840–45.

Simonson, Itamar and Emanuel Rosen (2014), *Absolute Value: What Really Influences*

Customers in the Age of (Nearly) Perfect Information, New York: HarperBusiness.

Sridhar, Shrihari and Raji Srinivasan (2012), "Social Influence Effects in Online Product

Ratings," *Journal of Marketing*, 76 (5), 70-88. <https://doi.org/10.1509/jm.10.0377>

Author Accepted Manuscript

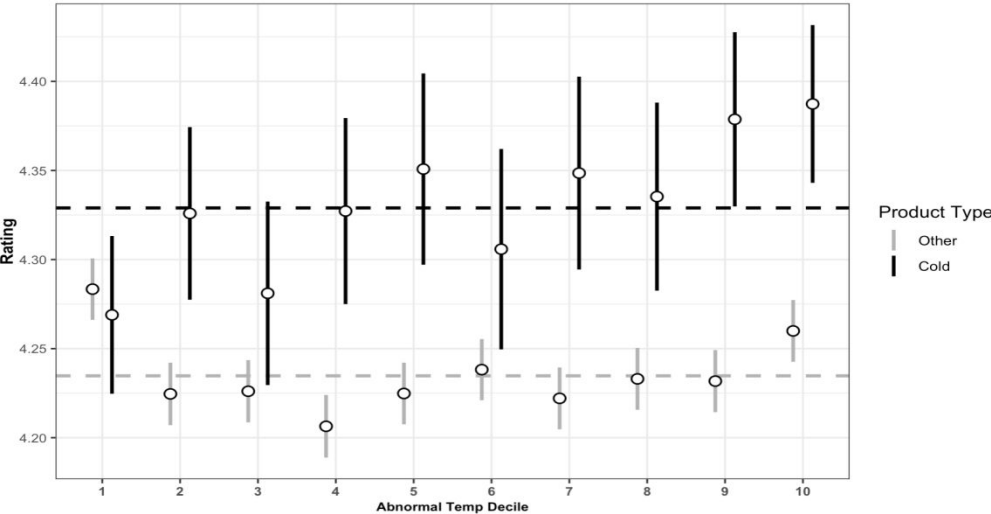
- Tirunillai, Seshadri and Gerard J. Tellis (2014), "Mining Marketing Meaning from Online Chatter Strategic Brand Analysis of Big Data using Latent Dirichlet Allocation," *Journal of Marketing Research*, 51, 463–79.
- Urbany, Joel E., Peter R. Dickson, and William L. Wilkie (1989), "Buyer uncertainty and information search," *Journal of Consumer Research*, 16 (2), 208–215.
- Ursu, Raluca M. (2018), "The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions," *Marketing Science*, 37 (4), 530–52.
- Varga, Marton and Paulo Albuquerque (2019), "Measuring the Impact of a Single Negative Consumer Review on Online Search and Purchase Decisions Through a Quasi-Natural Experiment," *Marketing Science Institute Research Report*, 19-03-01.
- Watson, Jared, Anastasiya Pocheptsova Ghosh, and Michael Trusov (2018), "Swayed by the Numbers: The Consequences of Displaying Product Review Attributes," *Journal of Marketing*, 82 (6), 109–131.
- Yin, Dezhi, Sabyasachi Mitra, and Han Zhang (2016), "When do consumers value positive vs. Negative reviews? An empirical investigation of confirmation bias in online word of mouth," *Information Systems Research*, 27 (1), 131–44.
- Zervas, Georgios, Davide Proserpio, and John W. Byers (2021), "A first look at online reputation on Airbnb, where every stay is above average," *Marketing Letters*, 32, 1–16.
- Zhu, Feng and Xiaoquan (Michael) Zhang (2010), "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics," *Journal of Marketing*, 74 (2), 133–48.

Author Accepted Manuscript
APPENDIX

Appendix A: Study 1 Binned Temperature Plot

To assess whether the effect we observe is similar across the range of temperatures, we binned each observation according to deciles of abnormal temperatures. Figure A1 presents the mean and 95% confidence interval of ratings for products within these bins, separated for cold-weather and other products. This illustrates a positive, linear effect for cold-weather gear, but a flat effect for other products. An exception is ratings for non-cold-weather products at the lowest decile of abnormal temperature, for which we have no theory-driven interpretation.

Figure A1: Product Ratings According to Deciles of Abnormal Temperatures



Note: Error bars correspond to 95% confidence intervals.

Author Accepted Manuscript

Quality in Context: Experience-Relevant Consumption Context Influences Product Ratings

Matt Meister

Assistant Professor of Marketing, University of San Francisco

2130 Fulton Street, San Francisco, CA 94117-1080

mmeister@usfca.edu

415-422-6721

Nicholas Reinholtz

Assistant Professor of Marketing, University of Colorado

995 Regent Drive, Boulder, CO 80309-0419

nicholas.s.reinholtz@colorado.edu

303-735-8019

Table of Contents:

Web Appendix	Contents	Page(s)
A	Details for Study 1	2–3
B	Bag-of-Words Sentiment Analysis for Study 1	4–6
C	Specification Curve Analysis for Study 2	7–8
D	Web Appendix Study 1: Replication of Study 3 in Winter Jackets	9–12
E	Web Appendix Study 2: Assessing Alternative Explanation for Study 3	13–15

These materials have been supplied by the authors to aid in the understanding of their paper. The AMA is sharing these materials at the request of the authors.

Web Appendix A:

Details for Study 1

Study 1 Density of Ratings and Temperatures

Figure W1. Density Histograms of Ratings Across Product Types

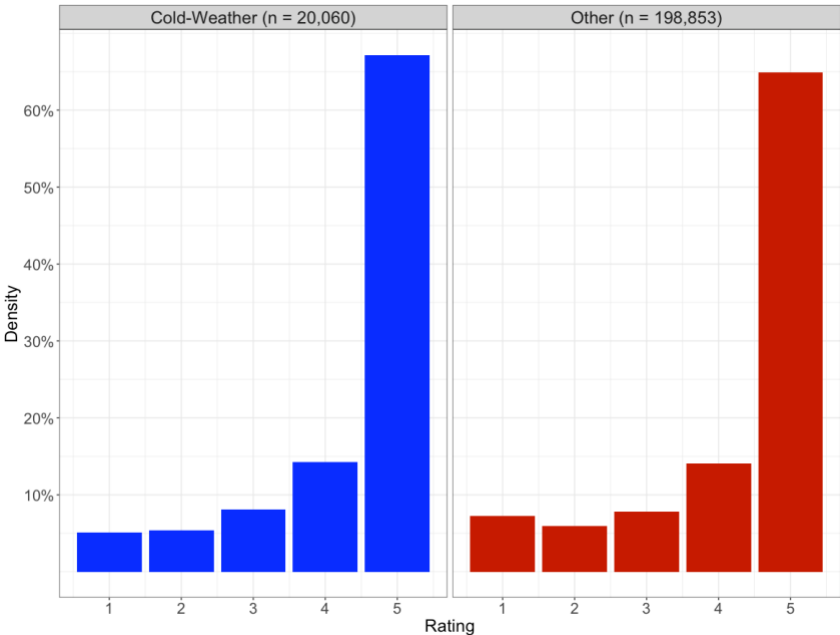


Figure W2. Density Plots of Absolute Temperatures Across Product Types

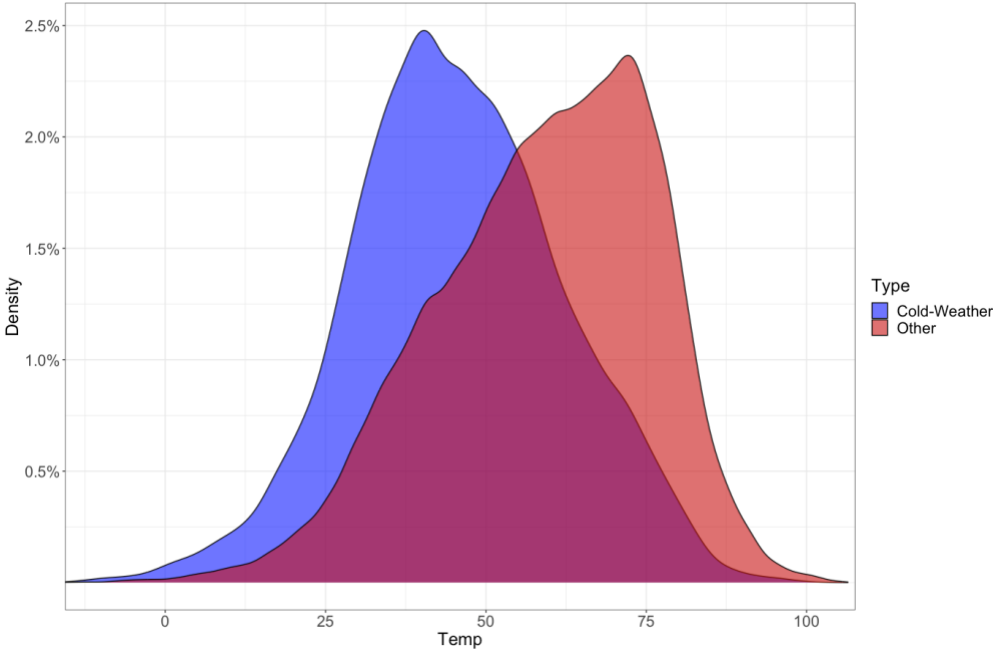
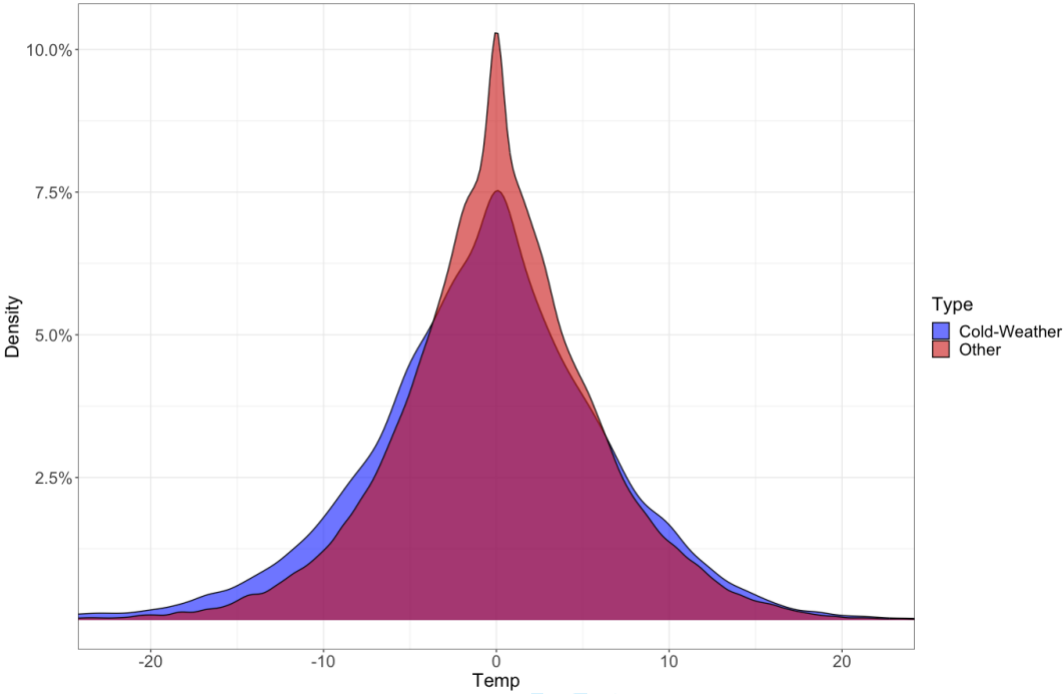


Figure W3. Density Plots of Location-Month De-meaned Temperatures Across Product Types



Web Appendix B:

Bag-of-Words Sentiment Analysis for Study 1

For each review, we first cleaned the raw text by removing punctuation and formatting, and substituting the most frequent negations (e.g., “no issues” was transformed to “perfect”). Then, we calculated the average sentiment of each review as the number of positive words minus the number of negative words, divided by the total number of positive and negative words. As a result, 4,946 reviews (2.3% of total) that had neither positive nor negative words were removed from this analysis, resulting in a final sample of 213,967 reviews. We multiplied sentiment by 100 for ease of presentation such that a value of 100 represents an exclusively positive review, – 100 exclusively negative, and zero neutral. The correlation between ratings and sentiment was high ($r = .469$), and the average review was relatively positive ($M_{Sentiment} = 56.329$), in line with the discrete ratings.

Replicating Table 1 with sentiment as the outcome demonstrates consistent results in every model (Table W1). Consistent with discrete ratings, we observe a significant interaction of product type and temperature in our most causally defensible model (Model 5; $\beta_{Interaction} = .102$, $t(182,323) = 3.746$, $p < .001$, 95% CI = [.054, .151], Std. $\beta = .036$).

Author Accepted Manuscript

Table W1: Regression Results with Sentiment as the Dependent Variable.

<i>Dependent Variable:</i>						
	(1)	(2)	(3)	(4)	(5)	(6)
Temp	-0.028	-0.072	0.051	-0.027	-0.016	-0.004
	(.008)	(.009)	(.011)	(.018)	(.018)	(.016)
CWG	-2.198	-2.869	-1.41	-2.925		
	(1.334)	(1.317)	(1.343)	(1.370)		
Temp x CWG	0.148	0.167	0.126	0.165	0.102	0.057
	(.023)	(.023)	(.023)	(.024)	(.025)	(.022)
Rating						18.361
						(.117)
Constant	4.269					
	(.014)					
Station FE	No	Yes	No	Yes	Yes	Yes
Month FE	No	No	Yes	Yes	Yes	Yes
Product FE	No	No	No	No	Yes	Yes
Observations	213,967	213,967	213,967	213,967	213,967	213,967
R ²	0.001	0.016	0.002	0.074	0.217	0.367
Adjusted R ²	0.001	0.006	0.002	0.005	0.081	0.257
Residual Std. Error	50.204	50.087	50.184	50.094	48.163	42.285
df	213,963	211,826	213,952	199,224	182,323	182,322

Note: "Temp" refers to the effect of one degree increase in the average daily mean temperature in the day of a review and two prior. "CWG" is an indicator that equals 1 if the product is cold-weather gear.

Author Accepted Manuscript

In Model 6, we add a control for rating to Model 5. In this model, the effect of context remains significant, indicating that—even after controlling for the numeric rating provided—review text is still affected consistently by context. However, this is likely due to measurement error in this case, as the correlation between sentiment and rating is much lower than that found with GPT in text. It does not appear that there is any hedging of negative or positive ratings, which would have resulted in a negative interaction coefficient in Model 6.

References:

Hu, Mingqing and Bing Liu (2004), “Mining and Summarizing Customer Reviews,” in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–77.

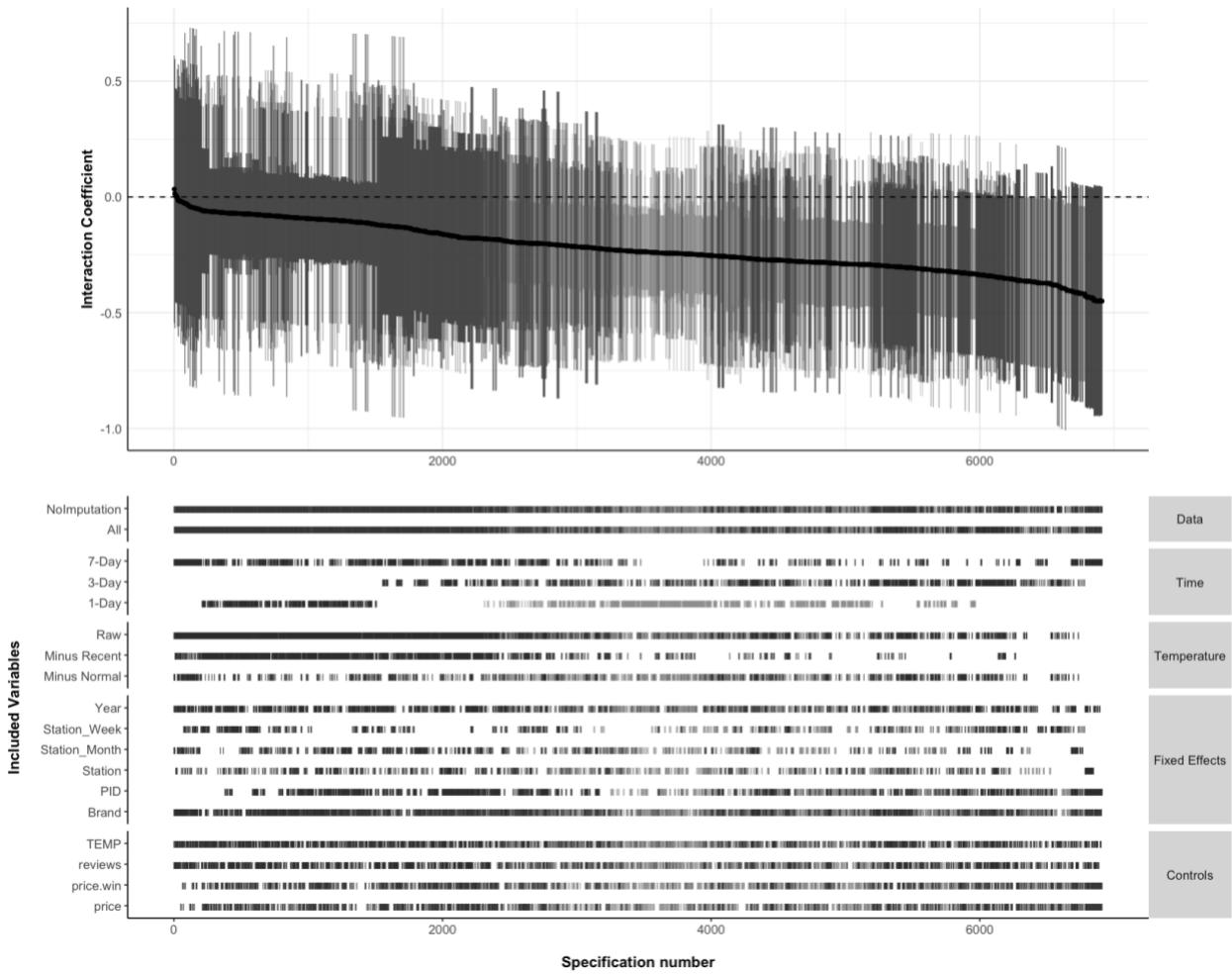
Author Accepted Manuscript

Web Appendix C:

Specification Curve Analysis for Study 2

We assessed robustness through a specification curve analysis, varying each combination of all potential analytic choices, yielding 6,912 models. These analytic choices are the same as in Study 1 with two exceptions. First, we only have three context measures (precipitation over either a one, three, or seven-day running average of a location's daily total) rather than the nine in Study 1. Second, rather than varying the inclusion of a control for precipitation, we vary the inclusion of a control for temperature. Results demonstrate negative interaction estimates in 6,890 (99.7%) of models, statistically significant in 1,792 (25.9%). Though this is less than the proportion of statistically significant models than in Study 1, this is likely due to the smaller sample of reviews for rain jackets, as the effect size is slightly larger in this study. Every model using the raw precipitation measure yielded a negative interaction effect, with 43.1% being statistically significant. Inconsistency in coefficient estimates is largely due to the use of different measures—models using observed minus recent precipitation as the independent variable led to coefficients closer to zero ($M = -.140$, $Q_1 = -.182$, $Q_3 = -.084$) than models using the raw measure ($M = -.301$, $Q_1 = -.353$, $Q_3 = -.255$) or precipitation minus normal ($M = -.229$, $Q_1 = -.283$, $Q_3 = -.193$).

Figure W4: Specification Curve for Study 2



Author Accepted Manuscript

Web Appendix D:

Web Appendix Study 1: Replication of Study 3 in Winter Jackets

Participants & Procedure

Eight-hundred and one participants were recruited from CloudResearch's approved Amazon Mechanical Turk participants and consented to complete this study. Of these, 778 completed the entire study, and 776 passed the attention check at the end of the study (which asked what product had been displayed). Consistent with our pre-registration, we removed these two participants who did not pass the attention check from our final sample.

After consenting to complete the study, participants were told to imagine that they had recently purchased a new winter jacket online. On the next page, participants were told to imagine that after the jacket arrived, they had used it as they normally would: *"Going to work, the grocery store, running errands, and walking outside."* Then, we randomly assigned participants to have had one of two experiences with the jacket. In the "as expected" condition, participants were told that *"While you wore this winter jacket, you noticed you felt as warm as you had hoped."* In the "worse" condition, participants were told that *"While you wore this winter jacket, you noticed you did not feel as warm as you had hoped."*

We also randomly assigned participants to one of three context information conditions. These conditions were intended to test the degree to which awareness could reduce context effects, similar to interventions used in the affect-as-information literature. We are agnostic as to which condition most closely replicates reality, as we cannot observe awareness in the REI data beyond its manifestation through mention in reviews. Instead, we draw attention to the attenuation between each, which measures the effect of *increasing* awareness.

In the “absent” condition, participants only read about their experience, and information about the context that led to that experience was absent. This simulates a consumer who is entirely unaware of their context. In the “light” reminder condition, participants saw information about the context that led to their rating. Specifically, in the warmer (colder) condition they were told that “*While you wore this winter jacket, the average temperature was slightly warmer (colder) than you would normally expect it to be at this time of year.*” This condition simulates a consumer who is highly aware of their context. Lastly, the “heavy” reminder condition simulates a consumer who is aware both of their context and its impact on experience. In this condition, participants saw the same information as in the other conditions, but also that “*This was probably the reason that you felt as warm (you did not feel as warm) as you had hoped when you bought the winter jacket.*”

After reading about their experience and context, the next page asked participants to rate the product on a 1–10 scale (1 = poor, 10 = excellent). In the light awareness condition, participants were shown the same context information from the page before. In the heavy awareness condition, we added the sentence “You should not let this temperature impact your rating” immediately before the rating. After rating their winter jacket, participants completed an attention check question asking what product they saw in the study, and were then paid.

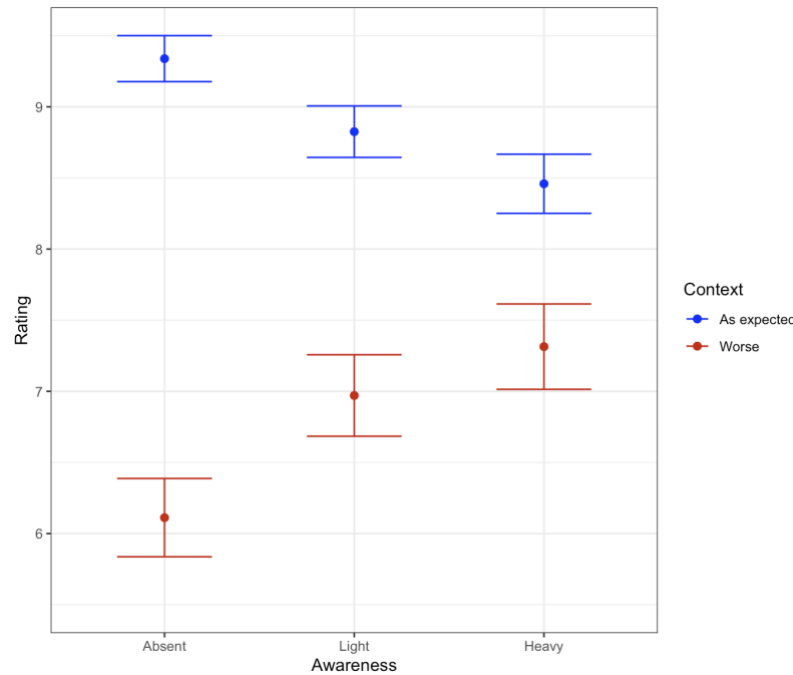
Analysis & Results

Our analysis of this study focuses on the attenuating effect of increasing awareness. Therefore, we analyzed ratings with linear regression, using ratings as the outcome, predicted by dummy codes for the light and heavy awareness conditions, a contrast code for context, and

Author Accepted Manuscript

interactions of context and reminder conditions. We included fixed effects for participants and standard errors were clustered by participants.

Figure W5: Results of Web Appendix Study 1



Note: Error bars represent 95% confidence intervals.

The simple effect of context in the absent awareness condition revealed significantly higher average ratings after “as expected” experiences than after worse experiences ($\beta_{Context} = 3.227$, $t(770) = 18.402$, $p < .001$, Std. $\beta = 1.792$). However, this effect was significantly attenuated by awareness of context (Figure W5). Specifically, we found a significant interaction of context and the light awareness condition ($\beta_{Context \times Light} = -1.172$, $t(770) = -5.554$, $p < .001$, Std. $\beta = -.762$) and of context and the heavy awareness condition ($\beta_{Context \times Heavy} = -2.082$, $t(770) = -8.375$, $p < .001$, Std. $\beta = -1.156$). Re-coding the model to compare the light and heavy awareness conditions found that the additional explanation of the impact of context on experience created significantly further attenuation beyond mere awareness we found a

Author Accepted Manuscript

significant interaction of context and the heavy (v. light) awareness condition ($\beta_{Context \times Heavy} =$
 $-.710, t(770) = -2.866, p = .004, \text{Std. } \beta = -.394$).

Peer Review Version

Author Accepted Manuscript

Web Appendix E:

Web Appendix Study 2: Assessing Alternative Explanation for Study 3

This study was designed to address the potential unrealistic nature of our Study 3. In real consumption, people are often aware of their contexts before consumption. For instance, people may check the weather in the newspaper before wearing their new winter coat, and almost certainly know their hunger level before eating at a restaurant. However, this order is flipped in our studies—people experience the context first, and then are made aware. It is possible that this order allows an anchoring-and-adjustment effect to explain our experimental findings, rather than correspondence bias. Specifically, that our experiment finds a lack of equivalence at strong reminders not because participants fail to completely attribute their experience to context, but because they anchor on their initial judgment. While this explanation is inconsistent with our field data, we address it directly here.

Participants & Procedure

Two-hundred and four participants were recruited from Prolific Academic and consented to complete this study.¹ This time, we pre-registered to exclude those who did not complete each question, which removed four participants. Of these, none provided more than one response, so our final data set consists of 200 participants.

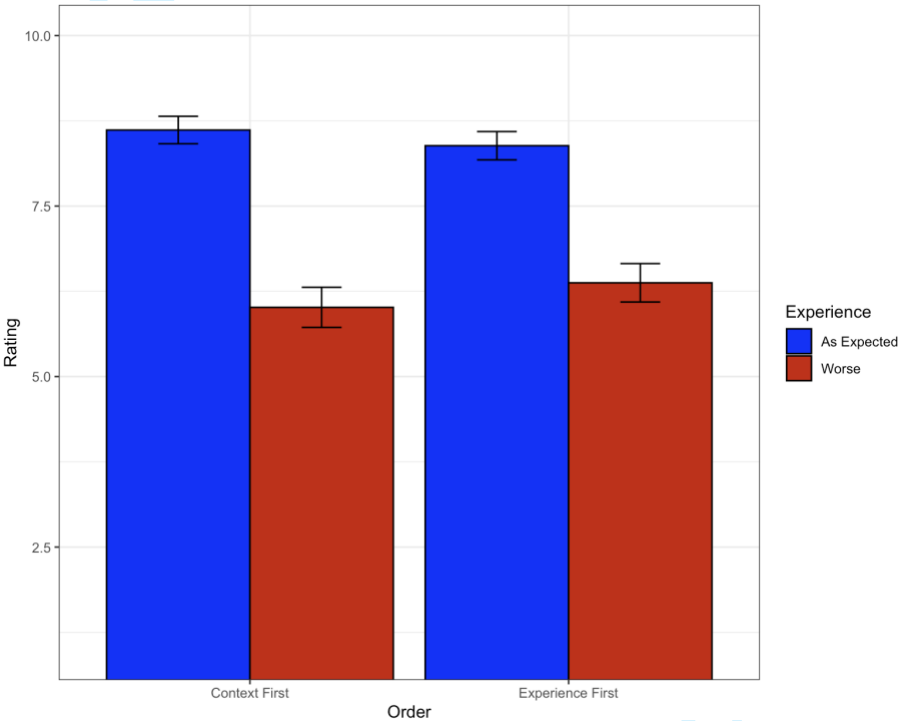
Participants in this study completed four product replicates (drawn without replacement for each participant from the six in Study 3). All participants saw the context information from the light context reminder condition. We varied whether the experience observed was as expected or not within-subjects, such that each saw two of each. This was fully crossed with the order in

¹ A consent question served as our attention check. To consent, participants had to click the answer furthest to the right on the screen.

which context reminders were presented, such that participants saw context information before the experience twice, and after the experience twice (once for a good experience, once for a bad). As with Study 3, participants rated each product on a 1–10 scale. Because each participant provided four responses, our final sample is of 800 ratings.

Analysis & Results

Figure W6: Results of Web Appendix Study 2



Note: Error bars represent 95% confidence intervals.

To investigate whether presenting context information first leads to more correction in ratings, we regressed ratings on order (dummy coded such that 1 indicates the experience coming first), experience, and their interaction. We included fixed effects for participant and product, and clustered standard errors by participant. This analysis revealed no main effect of order ($M_{\text{ContextFirst}} = 7.315$, $M_{\text{ExperienceFirst}} = 7.38$, $t(592) = 0.918$, $p = 0.36$, $d = -0.03$), a significant difference between experiences ($M_{\text{Worse}} = 6.195$, $M_{\text{As expected}} = 8.5$, $t(592) = 17.463$, $p < .001$, $d = -1.276$), and—unexpectedly—a significant interaction of experience and order ($\beta = -0.58$, $t(592)$

Author Accepted Manuscript

= -3.147, $p = 0.002$). Although we anticipated order having no effect, we instead found that participants provided ratings that were closer together for good and bad experiences if experience came first (Figure W6).

Discussion

Results from this study demonstrate that our findings from Study 3 were not due to the order in which we presented experiences and context in our paradigm. Rather, the effect observed in Study 3 was stronger when context came before the experience.