This article was downloaded by: [62.22.104.105] On: 03 July 2023, At: 04:12
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

When the Stars Shine Too Bright: The Influence of Multidimensional Ratings on Online Consumer Ratings

Christoph Schneider, Markus Weinmann, Peter N.C. Mohr, Jan vom Brocke

To cite this article:
Christoph Schneider, Markus Weinmann, Peter N.C. Mohr, Jan vom Brocke (2021) When the Stars Shine Too Bright: The Influence of Multidimensional Ratings on Online Consumer Ratings. Management Science 67(6):3871-3898. https://doi.org/10.1287/mnsc.2020.3654

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2020, The Author(s)

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 67, No. 6, June 2021, pp. 3871–3898
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

When the Stars Shine Too Bright: The Influence of Multidimensional Ratings on Online Consumer Ratings

Christoph Schneider,[a] **Markus Weinmann,**[b] **Peter N.C. Mohr,**[c,d] **Jan vom Brocke**[e]

[a] IESE Business School, University of Navarra, 08034 Barcelona, Spain; [b] Rotterdam School of Management, Erasmus University, 3062 PA Rotterdam, Netherlands; [c] Freie Universität Berlin, 14195 Berlin, Germany; [d] WZB Berlin Social Science Center, 10785 Berlin, Germany; [e] University of Liechtenstein, 9490 Vaduz, Liechtenstein

Contact: cschneider@iese.edu, https://orcid.org/0000-0003-2405-193X (CS); weinmann@rsm.nl, https://orcid.org/0000-0002-8342-2756 (MW); peter.mohr@neuroeconomics-laboratory.de, https://orcid.org/0000-0002-5179-060X (PNCM); jan.vom.brocke@uni.li, https://orcid.org/0000-0002-0071-3719 (JvomB)

Received: August 3, 2017
Revised: August 7, 2018; April 17, 2019; January 29, 2020
Accepted: March 18, 2020
Published Online in Articles in Advance: October 22, 2020

https://doi.org/10.1287/mnsc.2020.3654

Copyright: © 2020 The Author(s)

Abstract. Scholars generally assume that consumer ratings reflect consumer satisfaction, but ratings can be influenced by the design of the rating system. We examine two rating designs—single-dimensional rating systems, which elicit overall ratings only, and multidimensional (MD) rating systems, which elicit both dimensional and overall ratings—and how they impact overall ratings. Drawing on the accessibility–diagnosticity framework, we argue that dimensional ratings in MD systems influence overall ratings based on how the dimensions have been rated. We support this explanation with seven experiments. Our results suggest that across various experimental settings, rating objects, dimensions, and numbers of dimensions, overall ratings are systematically influenced by the design of the rating system.

History: Accepted by Chris Forman, information systems.

Open Access Statement: This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. You are free to download this work and share with others for any purpose, even commercially if you distribute your contributions under the same license as the original, and you must attribute this work as "*Management Science*. Copyright © 2020 The Author(s). https://doi.org/10.1287/mnsc.2020.3654, used under a Creative Commons Attribution License: https://creativecommons.org/licenses/by-sa/4.0/."

Funding: This work was supported by Liechtenstein Research Fund [Grant wi-18-1, wi-2-14], City University of Hong Kong [Grant 7004563 (IS)], Erasmus University Rotterdam, and IESE Business School.
Supplemental Material: The e-companion is available at https://doi.org/10.1287/mnsc.2020.3654.

Keywords: online ratings • biases • accessibility–diagnosticity framework • experiments

1. Introduction

Online ratings are often regarded as reflecting the aggregate of consumers' satisfaction (e.g., Ho et al. 2017), yet such ratings can be biased as a result of various factors, such as self-selection in consumer reporting (Li and Hitt 2008, Hu et al. 2009) or social influence (Moe and Trusov 2011, Muchnik et al. 2013, Wang et al. 2018).[1] Likewise, in line with behavioral economics and psychology research, which suggests that the design of the decision environment (i.e., the "choice architecture," Thaler et al. 2012) may influence people's choices, online ratings can be biased by the design of the rating system's user interface (Jiang and Guo 2015, Chen et al. 2018). These biases imply that ratings may not accurately reflect raters' level of satisfaction, potentially leading other consumers to make suboptimal decisions.

Against this background, we investigate the degree to which the design of a rating system influences consumer ratings and seek to determine the influence of information-systems design on overall ratings. More specifically, we examine two rating methods used on popular rating platforms and their impact on overall ratings: Single-dimensional (SD) rating systems, which ask consumers for an overall rating only, and multidimensional (MD) rating systems, which ask for detailed ratings about the product's attributes in addition to an overall rating. Thus, we seek to answer the following research question: *How do MD ratings influence overall ratings?* Drawing on theory from psychology—particularly the accessibility–diagnosticity framework—we propose that MD ratings make specific attributes of a product, service, or experience more accessible (i.e., the dimensions are more easily retrieved from a person's memory), thereby influencing overall ratings.

We test this proposition using seven studies containing a total of 17 conditions. We first conducted a series of three highly controlled studies with a focus on internal validity (Studies 1a–c). We then conducted Studies 2a–c to replicate and extend Study 1 using real-world experiences (rather than an artificial scenario).

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Finally, we conducted a field experiment (Study 3) to replicate Studies 1 and 2 in a realistic setting.

We demonstrate that asking consumers to provide dimensional ratings can systematically influence overall ratings in either a downward or upward manner. We suggest that the dimensional ratings (here operationalized by their mean) influence the overall rating because the dimensions become more accessible when a consumer forms an overall rating. Previous theoretical explanations have assumed that MD systems can lead only to upward deviations of overall ratings (Chen et al. 2018). We show that ratings can deviate in both directions: if the mean of the dimensional ratings is high, overall ratings tend to be higher than the overall rating from the SD system; conversely, if the mean of the dimensional ratings is low, overall ratings tend to be lower than the overall rating from the SD system. Further, we demonstrate that both the content and the number of dimensions to be rated influence overall ratings. Our main results—the influence of dimensional ratings on overall ratings—are robust across experimental settings, different rating objects (i.e., restaurants, short films, and universities), as well as content and number of dimensions.

By providing a better understanding of online rating systems, these findings have some important implications for three stakeholder groups: providers of rating systems, product/service providers, and consumers. In particular, we argue that providers of rating systems using dimensional ratings need to conduct large-scale testing to arrive at the mix of dimensions that best represents a product and may even consider using MD ratings only (without asking for overall ratings). Likewise, product/service providers interested in increasing sales need to carefully consider—and their products/services need to excel on—the dimensions included by rating systems in order to account for the effects of dimensional ratings on overall ratings. Finally, consumers are advised to avoid comparing ratings across systems because both the type of system and the dimensions rated can lead to significant differences in ratings; especially when evaluating products or services using MD ratings, it is important to be aware of how overall ratings are constructed and to ensure that the dimensions are aligned with one’s preferences.

2. Background

2.1. Online Ratings

Although e-commerce offers both sellers and consumers many benefits, one significant drawback is that consumers cannot experience and physically evaluate products or services before making a purchase decision, which increases uncertainty (Hong and Pavlou 2014). Because online reviews—typically including numerical ratings and/or textual descriptions—can reduce uncertainty (Bolton et al. 2004, Pavlou and Gefen 2004), consumers frequently seek recommendations from other consumers (Kwark et al. 2014) on sellers’ websites or on dedicated review platforms (such as the travel review site TripAdvisor).

Research has demonstrated that consumers use online ratings to forecast enjoyment (He and Bond 2013). Although online ratings may not necessarily correlate with objective measures of quality (de Langhe et al. 2016), they can serve as an important source of information for consumer decision-making (Simonson 2016), and research has consistently shown that reviews have a strong influence on sales (Dellarocas 2003, Godes and Mayzlin 2004, Chevalier and Mayzlin 2006, Liu 2006, Dellarocas et al. 2007, Duan et al. 2008, Forman et al. 2008, Chintagunta et al. 2010, Zhu and Zhang 2010, Archak et al. 2011, Ghose et al. 2012). Recognizing the strong effects of online reviews, companies respond strategically to reviews (Chen and Xie 2005, Dellarocas 2006, Hu et al. 2011)—for example, by encouraging users to provide them (Goes et al. 2014, Burtch et al. 2017)—which further indicates how important reviews have become for business success.

Online reviews—based on consumers’ postconsumption evaluations—are often considered more credible and less biased than descriptions provided by merchants because consumers (i.e., the reviewers) typically have no stake in the product (Bickart and Schindler 2001). Consumers perceive aggregated reviews as reflecting the “wisdom of the crowd” (Surowiecki 2004) and, consequently, also tend to perceive online shopping sites that present reviews as more useful than those that do not (Kumar and Benbasat 2006).

Whereas textual reviews can substitute to some extent for direct product experience, shoppers typically can read only a small subset of textual reviews—because of time constraints—and tend to focus on aggregated numerical ratings (de Langhe et al. 2016); these ratings, however, are subject to the influence of cognitive biases, which can undermine their helpfulness (Mudambi and Schuff 2010). In the next section, we review research on biases that have been shown to influence online rating behavior, particularly self-selection and reporting bias, social-influence bias, and dimensional-rating bias.

2.2. Biases in Online Ratings

2.2.1. Self-Selection and Reporting Bias. Researchers have shown that ratings do not necessarily represent

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

all consumers' aggregated evaluations of a certain product, which biases rating distributions in (at least) three ways. First, rating distributions tend to be skewed. For example, whereas existing positive evaluations tend to attract further positive evaluations, existing negative evaluations tend to attract fewer additional evaluations (Moe and Schweidel 2012); relatedly, consumers on eBay were shown to be more likely to post feedback when they were satisfied than when they were dissatisfied (Dellarocas and Wood 2008). Second, rating distributions tend to be bimodal because online star ratings are often biased toward the extreme positive or negative ends of the spectrum (Hu et al. 2009). For example, Hu and colleagues showed that the aggregated ratings of a particular product followed a normal distribution only in a controlled experiment; however, actual consumers' ratings of the same product on Amazon.com followed a J-shaped distribution with many one-star ratings and an even larger number of five-star ratings but comparatively few four-, three-, and two-star ratings. A commonly stated reason for J-shaped distributions is self-selection bias: Research has demonstrated that it is primarily consumers with highly favorable or highly unfavorable opinions who tend to devote effort to rating products, whereas people holding moderate opinions are less likely to do so (Hu et al. 2009, 2017). Third, rating distributions can be biased because ratings posted by early adopters tend to be higher (Li and Hitt 2008), whereas ratings posted by late adopters tend to be lower (Dellarocas et al. 2007, Godes and Silva 2012).

2.2.2. Social-Influence Bias.

Social-influence bias occurs when previously posted ratings influence subsequent ratings—often referred to as the "bandwagon effect" (Moe and Trusov 2011, Muchnik et al. 2013). The effects of social-influence bias, however, can be contradictory. On one hand, ratings can be biased positively because early adopters tend to be more positive in their ratings than late adopters (Li and Hitt 2008). On the other hand, ratings can be biased negatively because consumers are likely to recall negative aspects of a product (Sundaram et al. 1998, Hennig-Thurau et al. 2004) and because subsequent reviewers are just as likely to adjust their product ratings in a downward manner after observing negative ratings as they are to adjust them in an upward manner after seeing positive ratings (Schlosser 2005). In addition, friends' opinions can induce herding behavior, resulting in increased ratings among friends (Lee et al. 2015, Wang et al. 2018). Although the cumulative effects are unclear, these studies show that social influence can have a strong effect on consumer ratings.

2.2.3. Dimensional-Rating Bias.

The dimensional-rating bias suggests that a few dimensional ratings can influence people's overall ratings in MD systems. Whereas early research on rating biases primarily examined SD ratings (e.g., Li and Hitt 2008), some scholars have started to examine the effects of MD ratings (e.g., Tunc et al. 2017, Chen et al. 2018), often using text mining (e.g., Ghose et al. 2012, Ge and Li 2015). In particular, researchers have argued that MD ratings can mitigate the drawbacks and biases of SD ratings (described earlier) because they are superior in transferring consumer experiences (Archak et al. 2011, Godes and Silva 2012, Moe and Schweidel 2012, Chen et al. 2018). However, there is mixed evidence for this effect. Some scholars have argued that MD ratings improve information transfer, leading to higher overall ratings as expectations are better met (Chen et al. 2018); others have argued that overall ratings are skewed when consumers base their ratings on only a few, selectively accessible dimensions (Decker and Trusov 2010)—often toward the least satisfactory dimension (Liu et al. 2014)—an effect that is evident also in textual reviews (Ge and Li 2015). Together, these studies show that dimensional ratings affect consumer ratings; however, the effect is not yet fully understood.

In the context of restaurant ratings, Chen et al. (2018) investigated how others' dimensional ratings inform one's own ratings and found that overall ratings in MD rating systems (i.e., TripAdvisor) were significantly higher than overall ratings in SD rating systems (i.e., Yelp).[2] To explain the higher ratings of MD systems, Chen et al. (2018) proposed two mechanisms: information transfer and priming.

First, combining information-transfer theory and expectation-confirmation theory, they argued that the better information transfer of MD systems explains their higher overall ratings: Because MD systems present ratings on several dimensions to readers, these are able to form a detailed expectation prior to visiting a restaurant; more detailed expectations tend to better match actual experiences, leading to increased satisfaction, which would then be reflected in higher ratings. However, this mechanism assumes that all customers consult a rating platform before visiting a restaurant even though only 40% of U.S. adults do so (Smith and Anderson 2016). The expectations of the remaining customers—60%—are not influenced by previous ratings although many of them still provide a rating after visiting a restaurant. Moreover, an even smaller percentage is likely to view the dimensional ratings because, at the time of this writing, TripAdvisor presented dimensional ratings only on the individual entries of each restaurant (and not on the overview page).

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Hence, information-transfer theory and expectation-confirmation theory may at best only partially explain the observed effects.

Second, as an alternative explanation, Chen et al. (2018) suggest priming;[3] that is, the dimensional ratings may "prime" certain dimensions, which may then influence overall ratings. Having found no support for the priming explanation in their archival study, they designed an experiment (Chen et al. 2018, experiment 2) in which participants read four restaurant reviews of dining experiences and then rated the restaurant using either an SD or MD system. Although the ratings differed between the SD and MD systems, the differences were nonsignificant; consequently, Chen et al. (2018) concluded that the better information transfer of MD ratings—rather than priming—explains the higher ratings in MD systems (the information-transfer theory itself was not experimentally tested). However, participants were instructed to read only four rather short reviews (two to three sentences), significantly reducing their ability to judge the experience on multiple dimensions. Further, participants had to rate a restaurant after *reading reviews* of dining experiences rather than on the basis of *experiencing* the dining itself. It stands to reason that a judgment based on one's own experience is qualitatively different from a judgment based on the reading of a few short reviews that already contain others' judgments. Such mediated experiences are likely to differ from one's own lived experiences, which "tend to be stronger than 'second-hand' experiences in determining consumers' notion of reality" (Whelan and Wohlfeil 2006, p. 316); in particular, mediated experiences are likely to evoke much weaker reactions, reducing potential effect sizes. Moreover, providing a rating based on multiple reviews is akin to providing a summary of others' ratings rather than an evaluation based on one's own experience.[4] Finally, whereas the priming explanation is similar to the accessibility–diagnosticity framework that we draw on, Chen et al. (2018) made two additional restricting assumptions that may not hold true: (1) priming people with several dimensions in an MD rating system leads them to "consider all key aspects" (p. 4633), and (2) people are influenced by a negativity bias in SD ratings; consequently, using all dimensions should lead to a cancellation of the negativity bias and, thus, to more positive ratings regardless of which dimensions were included in the MD rating system. However, some studies have raised doubts about whether such negativity bias exists (see, e.g., Matlin and Stang 1978).

Here, we build on Feldman and Lynch's (1988) accessibility–diagnosticity framework and extend the work of Chen et al. (2018) by providing evidence for the existence of a dimensional rating bias. We argue that MD ratings may indeed influence overall ratings by making certain (but not necessarily all) dimensions more accessible. However, the mechanism on which we build our hypotheses differs in some important respects from the one proposed by Chen et al. (2018). Most importantly, in contrast to the explanation provided by Chen et al. (2018), the accessibility–diagnosticity framework suggests that dimensional ratings can lead not only to higher overall ratings (as Chen et al. (2018) argued), but also to lower overall ratings.

2.3. Order Effects in Evaluative Judgments

Order effects may play an important role when reviewers evaluate multiple dimensions. Since the early 1980s, psychologists and social scientists have studied response behavior in self-report surveys. Research in the field of survey methodology—in particular, the study of cognitive aspects of survey taking—has addressed a wide range of factors that may influence survey responses, such as how people's autobiographical memory affects retrospective reports of behaviors or how people make sense of a series of evaluative judgments (e.g., Jabine et al. 1984, Hippler et al. 1987). In particular, research on context effects in surveys—for example, so-called order effects—has repeatedly shown that preceding questions may influence people's responses to subsequent questions (see, e.g., Tourangeau and Rasinski 1988, Schwarz and Strack 1991, Peterson and Wilson 1992, Sudman et al. 1996, Schwarz 1999).

For example, Schwarz and Strack (1991) investigated order effects in questions about life satisfaction. Ratings of life satisfaction varied substantially depending on whether they were assessed before or after a question about marital satisfaction. When marital satisfaction was assessed after life satisfaction, ratings correlated modestly (0.18). However, when marital satisfaction was assessed before life satisfaction, ratings exhibited substantially higher correlations (0.67). Thus, answers about marital satisfaction exerted a much stronger influence on answers about life satisfaction when marital satisfaction was assessed earlier.

Several further studies have observed order effects in questions about even-handedness (for an overview, see Schuman and Ludwig 1982). Hyman and Sheatsley (1950), for example, asked their participants two questions related to freedom of the press: "Do you think a Communist country like Russia should let American newspaper reporters come in and send back to America the news as they see it?" and "Do you think the United States should let Communist newspaper reporters from other countries come in

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

here and send back to their papers the news as they see it?" Importantly, when the question about American newspaper reporters was asked first, 89.8% of the participants answered the question about Communist reporters in the affirmative; when the question about Communist reporters was asked first, only 36.5% answered this question in the affirmative. However, when each question was asked in the second position, approval rates substantially differed. Significantly fewer participants (65.6%) agreed with the statement about American newspaper reporters after they evaluated the one about Communist reporters, but approval rates for Communist reporters increased (73.1%), indicating that the answer to the second question was influenced by the presence of the first question.

These order effects and context effects can be attributed to the accessibility of relevant information: if people answer questions in a strict order, they use accessible information from preceding questions to answer subsequent ones. In an online rating context, these effects are likely to play a role when a reviewer rates multiple dimensions before providing an overall rating.

3. Theory and Hypotheses

3.1. Accessibility–Diagnosticity Framework

We argue that an MD rating system makes certain dimensions more accessible and that these dimensions consequently influence overall ratings. Our argument draws on accessibility–diagnosticity research (e.g., Srull and Wyer 1979, Tourangeau and Rasinski 1988, Schwarz and Strack 1991, Feldman 1992, Higgins and Brendl 1995, Schwarz 1998); in particular, we build on the work of Feldman and Lynch (1988), who argued that "momentarily activated cognitions have disproportionate influence over judgments made about an object or on related behaviors performed shortly after their activation" and that this "activation is a function of the environmental cues directing attention to some subset of the object's stimulus features, priming, retrieval factors, and individual differences" (p. 421). In particular, Feldman and Lynch (1988) summarized their accessibility–diagnosticity framework as follows:

> If the researcher asks a question about attitude (or belief, intention, etc.) and no such cognition already exists, the measure creates one. In answering subsequent questions in a survey, the newly created attitude may then be retrieved and used in generating answers to later questions about the same attitude object—creating, for example, a correlation between attitude and subsequently measured behavior. Had attitude not been measured prior to behavior, behavior would have been generated using different (nonattitudinal) inputs. Alternatively, consider those cases in which answers to all of the researcher's questions about beliefs, attitude, and so forth, already exist in long-term memory. The content of preceding questions in the survey, their ordering, and other aspects of measurement can still affect the observed relations among constructs. Instead of directly retrieving the appropriate response from long-term memory, subjects may retrieve some related response(s) made earlier in the survey. These earlier responses may be used directly to answer the question at hand, or integrated with other inputs to recompute an attitude or belief that already existed. (p. 431)

According to Feldman and Lynch (1988), the described effects depend on three main factors: (1) diagnosticity, (2) memory accessibility of potential inputs to judgments, and (3) accessibility of alternative inputs.

3.1.1. Diagnosticity.

> The perceived diagnosticity of the first judgment or decision for a second (later) one is the degree to which the respondent perceives that the answer to the first question correctly identifies how the second should be answered. A respondent's propensity to base an answer to the second question in a series on his or her answer to the first is a positive function of the perceived diagnosticity of the first question (Feldman and Lynch 1988, p. 424).

Feldman and Lynch (1988) argued that, when an answer to a previous question perceived as highly diagnostic is stored in working memory, it is unnecessary to compute an answer to the second question or to engage in an alternative retrieval strategy (which would result, for example, in a more effortful search of one's long-term memory). The aforementioned study by Hyman and Sheatsley (1950) provided an example of the effects of diagnosticity on judgments; the authors found order effects in questions about American and Communist newspaper reporters. Because both questions in their study referred to the general concept of freedom of the press, many respondents likely perceived the answer to the first question as highly diagnostic for the second one.

3.1.2. Memory Accessibility of Potential Inputs to Judgments.

> Memory accessibility, in turn, will be a function of the following conditions: (a) the time since the most recent activation of that cognition (Wyer & Srull 1986), (b) the amount of interfering material encountered in the same general content domain (Keller, 1987; Lynch et al. 1987), (c) elaboration and rehearsal of the original information (Ross, Lepper, Strack, & Steinmetz, 1977; Sherman, 1980; Sherman et al., 1978), (d) characteristics of the information itself that determine the rate of decay in the respondent's ability to retrieve it, such as vividness (Reyes, Thompson, & Bower, 1980) or abstraction and summarizing power (Chattopadhyay, 1986; Lingle, Geva, Ostrom, Leippe, & Baumgardner, 1979), (e) motivation and processing goals at the time

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

> of encoding of information (Biehal & Chakravarti, 1983; Loken & Hoverstadt, 1985), and (f) retrieval cues at encoding and retrieval, and so forth (Bettman & Sujan, in press). These conditions affect the likelihood that any previously formed cognition will be used as an input to a judgment—not just one activated in responding to previous questions in a survey. (Feldman and Lynch 1988, p. 426).

Feldman and Lynch (1988), thus, argued that, for information to be included in a judgment process, it needs to be not only diagnostic, but also accessible from memory. In turn, memory accessibility is influenced by many different factors. The likelihood that any particular piece of information is retrieved as an input to some judgment or behavior is, for example, inversely related to the amount of time since its most recent activation (Wyer and Srull 1986) and the amount of information belonging to the same content domain processed in the interim (Keller 1987). Furthermore, Wyer and Srull (1986) hypothesize that people first search their working memory when being asked to provide a judgment. However, working memory capacity is limited, decreasing the probability that a specific piece of information is (still) stored after time has elapsed and novel pieces of information belonging to the same content domain have started occupying working memory. Wyer and Srull (1986), therefore, argued that the search terminates if a sufficient basis for making the judgment is found. Information that is not (any longer) stored in working memory might, therefore, not enter the judgment process. Earlier judgments might, thus, affect later ones because information that was recalled while answering the earlier question has a high likelihood of being recalled from working memory when answering a later (summary) question. Consequently, the effect of early judgments on later ones might be at its strongest when they occur in direct succession.

A number of studies have also indicated that elaboration on or rehearsal of a piece of information increases the likelihood that it will be retrieved for use in making a later judgment. Sherman et al. (1978), for example, found that people who recalled how they rated the importance of recycling behaved consistently with their ratings. However, people who were not induced to elaborate on their ratings did not. Relatedly, Higgins (1978) showed that causing people to elaborate on favorable (unfavorable) information about a specific person led to more positive (negative) evaluations of that person. Earlier questions in a survey belonging to the same content domain might, thus, work in the same way, causing people to (subconsciously) elaborate on the specifics asked in the question.

Another important factor influencing memory accessibility is the presence of retrieval cues. In a seminal study, Bettman and Sujan (1987) primed participants with words related to a specific decision criterion (i.e., creativity or reliability). For the priming manipulation, the authors used a separate, unrelated task (framed as a pretest for a word perception test). In this task, words likely worked as retrieval cues, increasing the accessibility of pieces of information consistent with the cue. After the priming, participants (1) had more thoughts related to that criterion, (2) gave higher ratings to the importance of attributes related to that criterion, and (3) had higher preference ratings for products that were described with respect to that criterion (in contrast to the other one). In line with these findings, earlier questions (or later ones if people can go back and forth) might work as retrieval cues that prime people to use specific pieces of information during the judgment process of answering the question at hand.

3.1.3. Accessibility of Alternative Inputs.

> The increased accessibility of an input . . . simultaneously reduces the likelihood that other inputs will be retrieved from memory because of output interference effects. (Feldman and Lynch 1988, p. 428)

Feldman and Lynch (1988) argued not only that information related to earlier questions might become *more* accessible, but also that other pieces of information might become *less* accessible. A likely reason for this effect is the limited capacity of working memory: if specific pieces of information enter working memory because they were recently activated, others are no longer stored and, therefore, become less accessible. In other words, Feldman and Lynch (1988) argued that people base their responses only on the subset of relevant cognitions that are most accessible in memory (such as responses to a prior question) even if other (less accessible) cognitions or responses were in long-term memory. Consequently, people are less likely to use older pieces of information in a judgment process.

3.2. Hypotheses

In the context of online ratings, we rigorously apply the accessibility–diagnosticity framework proposed by Feldman and Lynch (1988) to make specific predictions about how overall ratings might differ between SD and MD systems. In particular, we focus on the memory-accessibility part of the framework to explain why overall ratings differ between SD and MD systems.[5]

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

In the SD rating process, people have to provide only an overall rating, so they search their memory for relevant information. If they have already formed an overall evaluation (e.g., directly following an experience), they likely recall this information and, if necessary, convert it into a numeric value of the rating scale. If no such overall evaluation is stored in memory, people compute an answer by recalling and (if necessary) evaluating experiences stored in memory. As Chen et al. (2018) argue, negativity bias may influence such a computation (Kanouse and Hanson 1987) because negative experiences have a higher likelihood of being stored in memory compared with positive ones (e.g., Baumeister et al. 2001; for evidence against this negativity bias, see Matlin and Stang 1978).

In the MD rating process, people are asked to rate experiences on several dimensions before (or after) providing an overall rating. Again, we can distinguish two possible cases: either the individual has already formed an overall evaluation (e.g., directly following an experience) or the individual has to compute it on the spot. (1) If an individual has already formed an overall evaluation—in other words, the evaluation lies in long-term memory—the individual may use responses to dimensional ratings stored in working memory directly to provide an overall rating or may integrate those responses with other memories to recompute the overall evaluation; because the responses to the dimensional ratings are highly accessible (as highlighted previously), the initial rating (from long-term memory and, hence, less accessible) is likely to factor less into the overall evaluation. (2) If no overall evaluation is stored in memory, the overall rating has to be computed. In this case, people rely mainly on information they can readily recall from working memory and stop searching when they have retrieved enough information to arrive at an overall rating. Because information related to dimensional ratings was recently activated, it is likely highly accessible; overall ratings, therefore, strongly rely on these pieces of information. In both cases (the individual has already computed an overall evaluation or must compute one in order to provide an overall rating), information related to dimensional ratings is more likely to impact the overall rating compared with situations in which *only* an overall rating has to be provided.

When people are coming up with an overall rating, dimensional ratings can, thus, lead them to consider information they might not have retrieved otherwise but that they might now integrate into the evaluation process. The overall rating is then influenced by the valence of the information cued by the dimensional ratings—either positive, neutral, or negative on average. If dimensional ratings tend to be positive on average, the overall rating tends to be more positive; conversely, if dimensional ratings tend to be negative on average, the overall rating tends to be more negative.[6] If dimensional ratings are on average rather neutral, one might expect no effect on overall ratings. Combining these arguments, we propose the following:

Proposition 1. *Overall ratings in SD and MD systems are driven by the accessibility of relevant information, such that, in MD systems, dimensional ratings increase the accessibility of information related to the dimensions.*

In the following paragraphs, we explore the potential consequences of the proposed effect and present hypotheses that can be empirically tested. We proposed that MD systems increase the likelihood that evaluations from dimensional ratings enter into the overall rating. Consequently, we expect that the overall rating in an MD system is influenced by the dimensional ratings, such that the overall rating tends toward the mean of the dimensional ratings. In other words, depending on the dimensions asked, overall ratings from MD and SD systems are likely to differ because the overall rating from the MD system is influenced by its dimensional ratings.

To illustrate the proposed effects, we provide five scenarios about a fictional character, Mary, visiting a restaurant and rating her experiences on either an SD or MD system to explain how dimensional ratings influence overall ratings. Thereby, we focus on examples in which Mary has not stored an overall evaluation in memory but must compute one before providing an overall rating. Please note that the following examples are hypothetical and stylized, meaning especially that the number of experiences considered might differ between individuals. However, the examples are meant to illustrate that the likelihood of a person using specific pieces of information increases if the acessibility of the information increases, thereby influencing overall ratings in the direction of the mean of the dimensional ratings.

3.2.1. Scenario 1 (SD System; Baseline).

As a baseline, let's assume an SD system in which Mary bases her overall rating on four experiences (food, service, value, and atmosphere). (Note that an SD system does not ask for dimensional ratings though we assume here that these are the underlying experiences that influence Mary's overall rating.) In her mind, Mary would subconsciously rate these experiences as follows: 1 star (out of 5) for service, 2 stars for food, 2 stars for value, and 5 stars for atmosphere. Averaging these numbers, Mary's SD overall rating would be 2.5 stars (for simplicity, we assume equal weight across all experiences). This scenario is meant to serve as the

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

baseline for comparing the following MD scenarios (see Table 1, scenario 1).

3.2.2. Scenario 2 (Dimensional Ratings Exert an Upward Influence on Overall Ratings). Second, given the same restaurant, let's assume an MD system in which Mary has to rate two dimensions before providing an overall rating: the previously mentioned experience food (which she would have rated 2 stars) and the atmosphere (which she would have rated 5 stars). The mean of these dimensional ratings is, thus, $(2 + 5)/2 =$ 3.5 stars. Consequently, her overall rating—which would have been 2.5 stars in the SD system—would tend upward in the direction of the dimensional rating's mean (3.5 stars) (see Table 1, scenario 2).

3.2.3. Scenario 3 (Dimensional Ratings Exert a Downward Influence on Overall Ratings). Third, this process can go in both directions: upward if the dimensional ratings are comparatively high and downward if the dimensional ratings are comparatively low. Assume that the MD system asks Mary to rate an experience she would have rated 2 stars in the SD scenario (i.e., food) and another experience she would have rated 1 star (i.e., service). The mean of these dimensional ratings is, thus, $(2 + 1)/2 = 1.5$ stars. As a result, her MD overall rating would tend downward in the direction of the dimensional rating's mean (1.5 stars) (see Table 1, scenario 3).

3.2.4. Scenario 4 (Dimensional Ratings Exert No Influence on Overall Ratings if Dimensional Ratings Are Assessed After Overall Ratings). Fourth, we expect differences in overall ratings between SD and MD systems only if dimensional ratings were assessed prior to overall ratings. Assume that Mary would have to provide an overall rating *before* rating the previously mentioned experiences food (2 stars) and atmosphere (5 stars). In this case, the MD overall rating should not differ from the SD overall rating (i.e., both 2.5 stars). However, the (prior) overall rating might influence the (subsequent) dimensional ratings. In contrast to the situation in which dimensional ratings were assessed first and the mean of these dimensional ratings was $(2 + 5)/2 = 3.5$ stars, this mean might now tend downward in the direction of the MD overall rating.

3.2.5. Scenario 5 (Dimensional Ratings Exert No Influence on Overall Ratings if Many Dimensions Are Highlighted). Fifth, we expect differences in overall ratings between SD and MD systems to diminish as more dimensions are highlighted. Whereas few dimensions might cover only some but likely not all relevant aspects and, therefore, bias ratings, an increasing number of dimensions would likely lead to a more complete coverage of relevant aspects and, therefore, reduce the rating bias. Assume that Mary would have to rate not only the previously mentioned experiences food (2 stars) and atmosphere (5 stars) before providing an overall rating (as in Scenario 2) but also value (2 stars) and service (1 star). The mean of these dimensional ratings, $(5 + 2 + 2 + 1)/4 = 2.5$, is, thus, similar to the SD overall rating in Scenario 1. Whereas we expected a difference in Scenario 2, this difference would disappear because further dimensions are added to the rating.[7]

In summary, given our theoretical explanation and the examples presented, we hypothesize that order effects in online ratings will occur. Again, please note that the following hypotheses assume that the assessed dimensions are high in diagnosticity (as is usually the case in the practice of online ratings). The hypothesized effects might not necessarily hold in case of low diagnosticity.

Hypothesis 1. *MD overall ratings will be higher (lower) than SD overall ratings if dimensional ratings are assessed before overall ratings and the average of the dimensional ratings is higher (lower) than SD overall ratings.*

Hypothesis 2. *MD overall ratings will not differ from SD overall ratings if dimensional ratings are assessed after overall ratings.*

Table 1. Summary of Five Hypothetical Scenarios Illustrating How Dimensional Ratings Influence Overall Ratings (OR)

Scenario	Description	Se	Fo	Va	At	Mean	Rating deviation
1	SD (recalling experience)	1	2	2	5	2.5	–
2	MD (highlighting positive dimensions before OR)		2		5	3.5	Upward
3	MD (highlighting negative dimensions before OR)	1	2			1.5	Downward
4	MD (highlighting positive dimensions after OR)		2		5	3.5	Stable
5	MD (highlighting many dimensions before OR)	1	2	2	5	2.5	Stable

Notes. Se = service; Fo = food; Va = value; At = atmosphere. "Rating deviation" describes the deviation of the overall rating in an MD system from that in an SD system.

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Hypothesis 3. *The dimensional rating bias will decrease as the number of dimensional ratings assessed increases such that the difference between MD overall ratings and SD overall ratings decreases.*

4. Overview of Experiments

We conducted a series of between-group design experiments—participants were randomly assigned to either SD or different versions of MD conditions—to test our hypotheses. To rule out the information-transfer explanation as well as other possible (social) influences, participants had neither information on prior ratings (we provided no aggregated score) nor access to textual reviews; they consequently had no expectations that needed to be met (to control for the information-transfer effects proposed by Chen et al. (2018)) and no social influence that could bias their ratings (Muchnik et al. 2013). Whereas Study 1 was a series of three highly controlled experiments with a focus on internal validity, Studies 2 and 3 focused on ecological validity.[8] In particular, the experiments in Study 1 used a controlled scenario in the context of restaurant reviews to maximize internal validity. In Study 2, we replicated and extended Study 1 using a series of experiments with lived experiences (rather than an artificial scenario) in the context of movie ratings. In Study 3, a field experiment, we replicated Studies 1 and 2 in a realistic setting, asking students to rate their own university (i.e., a lived experience of longer duration and higher involvement). The materials and instructions are provided in e-companion section EC.2; data and analysis scripts are provided at https://osf.io/sw8ax.[9] Table 2 provides an overview of the studies, their purposes, and our findings; Table 3 provides an overview of the different conditions used; and Table 4 compares the results of our regression analyses.[10] Together, these results support our proposed mechanism and suggest that dimensional ratings make their evaluation more accessible, thereby influencing overall evaluations.

5. Study 1: Experiments on Restaurant Ratings

5.1. Overview

We conducted a series of experiments in a controlled setting in the context of restaurant ratings to test our hypotheses, that is, whether dimensional ratings bias overall rating (1) when we ask for dimensional ratings *before* overall ratings (Hypothesis 1, Study 1a), (2) when we ask for dimensional ratings *after* overall ratings (Hypothesis 2, Study 1b), and (3) when we increase the number of dimensions assessed prior to an overall rating (Hypothesis 3, Study 1c). In addition, we explored whether merely considering dimensions—without rating them—influences overall ratings (Study 1b). Next, we describe the elements that were common across the experiments on restaurant ratings; we then present the specific details of each study before providing a general discussion of the findings.

5.1.1. Procedure and Materials. Studies 1a–c followed the same general procedure. At the beginning of each experiment, we asked participants to read a scenario of a restaurant visit (we did not inform the participants that the visit would have to be rated; the scenario is described as follows and in more detail in e-companion section EC.2). After reading the scenario, participants watched a short film; this served to

Table 2. Overview of Experiments

Study	Purpose	Sample	Findings
Experiments on restaurant ratings			
1a	Test Hypothesis 1 a controlled environment by manipulating rating method (MD versus SD)	166 Prolific workers	MD ratings significantly influence overall ratings
1b	(1) Test Hypothesis 2—whether MD ratings influence overall ratings even if elicited *after* overall ratings (2) Test whether merely considering dimensions—without rating them—influences overall ratings	826 Prolific workers	(1) MD ratings significantly influence overall ratings even if elicited after overall ratings (2) Merely considering dimensions—without rating them—influences overall ratings
1c	Test Hypothesis 3—whether the number of dimensions affects the overall rating	392 Prolific workers	If many dimensions are highlighted, there is no significant effect of MD ratings
Experiments on movie ratings			
2a	Replicate Study 1a and test Hypothesis 1 using an immediate experience	192 Prolific workers	MD ratings did not significantly influence overall ratings
2b	Test Hypothesis 1 (MD ratings may bias overall ratings in a downward manner)	198 Prolific workers	MD ratings exert a significant downward influence on overall ratings when mean MD ratings are low
2c	Test the effect of varying the content of the dimensions (keeping number of dimensions constant)	847 Prolific workers	The effect of MD ratings on overall ratings is dependent on the content of the dimensions
Field experiment on university ratings			
3	Test Hypothesis 1 in a real-life situation using a field experiment	142 students	MD ratings significantly influence overall ratings

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Table 3. Overview of Conditions Across Experiments

	Conditions							
Study	SD	MD (before SD)	MD (after SD)	MD (priming)	MD (omit negative dimension)	MD (omit positive dimension)	MD (5 dimensions)	MD (10 dimensions)
Experiments on restaurant ratings								
1a	×	×						
1b	×	×	×	×				
1c	×						×	×
Experiments on movie ratings								
2a	×	×						
2b	×	×						
2c					×	×		
Field experiment on university ratings								
3	×	×						

introduce a distraction and a time lag between the "experience" and the rating as would typically be the case in real-world situations.[11]

After watching the short film, participants rated the restaurant visit by providing an overall score and/or by rating the restaurant visit on the rating dimensions *food*, *service*, *value*, and *atmosphere* (which were the dimensions used by TripAdvisor at the time the experiment was conducted);[12] the details of this experimental manipulation differed depending on the objective of the study and are described separately. Finally, the participants completed a demographic survey. The sessions lasted an average of 10 minutes. All materials were presented in English using the Qualtrics survey platform.

We developed the fictitious restaurant-visit scenario based on the DINESERV instrument, a 29-item scale for measuring consumer perception of service quality in restaurants (Stevens et al. 1995) that is based on the SERVQUAL instrument (Parasuraman et al. 1988). DINESERV consists of five dimensions: *reliability*, *assurance*, *responsiveness*, *tangibles*, and *empathy*. We framed four dimensions (reliability, assurance, responsiveness, and empathy) as positive and one dimension (tangibles) as negative; this served to increase variance and allowed us to confirm whether participants read the scenario carefully. (See e-companion section EC.2 for the scenario-development process.) We pretested the scenario to ensure that it would be understood by our participants. In particular, we first discussed the scenario with a focus group at one author's university to ensure the clarity of the wording. Further, we conducted four pretests using a total of 347 participants recruited from the online recruiting platform Prolific (https://prolific.co) to test various aspects of our study; in all pretests, participants consistently identified the positive and negative dimensions on the DINESERV dimensions correctly, suggesting that participants were highly likely to understand the scenario correctly (see e-companion section EC.3 for the results).

5.1.2. Participants. We calculated the required sample size for each study using G*Power (Faul et al. 2009).[13] We used Prolific (https://prolific.co) to recruit participants who met the following criteria: reside in English-speaking countries (Australia, Canada, England, New Zealand, and the United States), be at least 18 years old, have an approval rating of at least 90% for previously completed studies, and have not participated in any of our other studies. We paid all participants £1 (about US$1.25) for a 10-minute task, a payment equivalent to an hourly wage of £6 (US$7.50). We used Qualtrics' Randomizer to randomly assign participants to the conditions. Given the potential for automated, malicious, or random responses on microtask marketplaces such as Prolific or Amazon mTurk (see Kittur et al. 2008), we redirected participants who spent less than 10 seconds on the scenario page to the end of the study and did not provide them with any compensation and removed any incomplete cases. In view of the possibility that not all participants carefully read the scenario, we include only those participants who spent more than 50 seconds reading the scenario page in our analyses.[14] Table 5 provides details about the participant demographics. Table 6 provides details about the number of participants in the different conditions.

5.1.3. Measures. We captured the rating scores on a nine-star scale. To compare the rating scores from various conditions, we created three dummy variables (separately described later): $SD_{overall}$ is the overall rating from the SD condition, $MD_{overall}$ is the overall rating from the MD condition, and $MD_{dimensional}$ is the mean of the dimensional ratings from the MD condition, that is, the mean rating of the dimensions *food*, *service*, *value*, and *atmosphere*.[15] Thus, we

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Table 4. Mixed-Effects Regression Results of Experiments

DV: Rating	(1a) Restaurants (MD versus SD)	(1b) Restaurant (order effects)	(1c) Restaurant (number of dimensions)	(2a) Movie (positive influence)	(2b) Movie (negative influence)	(3) University
Fixed						
(1) Intercept (SD rating)	6.23*** [6.00, 6.45]	6.33*** [6.18, 6.48]	6.44*** [6.25, 6.63]	5.87*** [5.50, 6.24]	5.72*** [5.29, 6.15]	6.15*** [5.78, 6.53]
(2) MD-first (overall rating)	0.52*** [0.21, 0.83]	0.27* [0.06, 0.48]		0.15 [−0.36, 0.66]	−1.03*** [−1.62, −0.44]	0.72** [0.21, 1.22]
(3) MD-first (mean dimensional rating)	0.94*** [0.63, 1.26]	0.52*** [0.31, 0.73]		0.81*** [0.30, 1.32]	-0.67* [−1.26, −0.08]	0.50† [−0.01, 1.00]
(4) MD-last (overall rating)		0.19† [−0.01, 0.40]				
(5) MD-last (mean dimensional rating)		0.75*** [0.54, 0.96]				
(6) MD-prime (overall rating)		0.18† [−0.03, 0.39]				
(7) MD-Top5 (overall rating)			0.40** [0.12, 0.67]			
(8) MD-Top5 (mean dimensional rating)			0.69*** [0.42, 96]			
(9) MD-Mixed10 (overall rating)			0.14 [−0.13, 0.40]			
(10) MD-Mixed10 (mean dimensional rating)			0.20 [−0.06, 0.47]			
Random						
σ^2	0.33	0.53	0.42	0.63	0.67	0.41
$\tau_{00,level\ 2}$	0.72	0.64	0.78	2.57	3.78	1.93
$ICC_{level\ 2}$	0.69	0.55	0.65	0.80	0.85	0.82
$N_{level\ 2}$	166	826	392	192	198	142
$Obs._{level\ 1}$	253	1,239	661	293	304	219
Marginal R^2	0.12	0.05	0.04	0.04	0.04	0.04
Conditional R^2	0.72	0.57	0.66	0.81	0.86	0.83

Notes. The table compares the results of the various experiments. The intercept represents the overall rating from the single-dimensional rating condition (SD); all other variables/conditions represent deviations of the respective condition from the SD rating. CI in brackets.

† $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$.

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Table 5. Study 1: Participant Demographics

	Study 1a	Study 1b	Study 1c
Participants (recruited)	200	1,005	600
Participants (analyzed)	166	826	392
Average age	35.5	34.0	35.0
Women, %	56.6	53.6	50.5
Average reading time (in seconds)	131.0 (251.4)	119.2 (113.9)	126.4 (113.0)

Note. SD in parentheses.

obtained one rating score for participants in the SD condition and two rating scores for participants in the MD conditions.

5.1.4. Model Specification. In studies 1a–c we had multiple rating scores per participant in the MD conditions (i.e., participants had to provide both an overall rating and dimensional ratings); thus, we had to consider that observations from the same participant might be correlated (Gelman and Hill 2007). To account for the clustered nature of our data—the observations (level 1) are nested within the participants (level 2)—we used a linear mixed-effects regression model. Such models account for between-subject variability by allowing individual intercepts to vary (i.e., a varying intercept for *participants*—in other words, a random effect).[16] Thus, we specified a linear mixed-effects regression model (with a varying intercept for *participants*) to compare the rating scores from different conditions, in which the overall rating in the SD condition serves as the baseline. Here, we present the model for Study 1a as an example:

$$\begin{aligned} \textit{rating score}_{ij} = \beta_0 + \beta_1 \cdot MD_{overall_{ij}} \\ + \beta_2 \cdot MD_{dimensional_{ij}} + u_{0j} + \epsilon_{ij}, \end{aligned} \tag{1}$$

where i indicates ratings from j participants and β_0 represents the mean rating in the SD condition. The other beta values (β_1 and β_2) represent each condition's deviation from the mean of the SD condition: β_1 is the difference between the overall rating provided in the SD condition and that provided in the MD condition; β_2 is the difference between the overall rating provided in the SD condition and the mean dimensional rating in the MD condition—the hypothesized effect of accessibility. The term u_{0j} is a level-two random effect (i.e., on the participant level) that describes the between-subject variability of the outcome variable *rating score* and captures the nonindependence between observations i from the same participant j; further, u_{0j} is assumed to be normally distributed with mean zero and constant variance. The term ϵ_{ij} indicates level-1 residuals (i.e., on the observation level), which are assumed to be normally distributed with mean zero and constant variance. To estimate the mixed-effects models, we used the statistical software package *R* (Ihaka and Gentleman 1996) with the *lme4* package (Bates et al. 2015)—in particular the *lmer* function.[17]

5.2. Study 1a: Effect of Dimensional Ratings

The purpose of Study 1a was to test Hypothesis 1—whether eliciting dimensional ratings can influence overall ratings—in a highly controlled setting. We randomly assigned participants to one of two conditions: an overall-rating-only condition (SD condition) and a dimensional-rating condition with a subsequent overall rating (MD condition).

5.2.1. Procedure and Measures. Participants in the SD condition provided an overall rating only ($\beta_0 = SD_{overall}$, which served as the baseline rating). Participants in the MD condition provided dimensional ratings immediately before providing an overall rating (on the same page); thus, we obtained two ratings: an overall rating ($\beta_1 = MD_{\text{overall}}$) and a mean dimensional rating—that is, the mean rating of the dimensions *food*, *service*, *value*, and *atmosphere* ($\beta_2 = MD_{dimensional}$). We followed the model specification presented in Equation (1).

5.2.2. Results. The purpose of Study 1a was to test our hypothesis regarding the effect of asking for dimensional ratings *before* an overall rating (see Table 7 for summary statistics). Because differences between overall ratings are expected only if dimensional ratings deviate from the SD overall rating, we first investigated whether the mean dimensional ratings in the MD condition differed from the overall ratings in the SD condition; we found that the mean dimensional ratings in the MD condition were significantly higher than the overall ratings in the SD condition (7.17 stars versus 6.23 stars, $\beta_2 = .94$, $p < .001$). Having confirmed that differences in overall ratings were to be expected, we proceeded with testing our hypothesis and found that overall ratings in the MD condition were also significantly higher than those in the SD condition (6.75 stars versus 6.23 stars, $\beta_1 = .52$, $p < .001$). These results support Hypothesis 1,

Table 6. Study 1: Participants per Condition

	Study 1a	Study 1b	Study 1c
Participants (analyzed)	166	826	392
SD	79	206	123
MD	87		
MD-first		203	
MD-last		210	
MD-prime		207	
Top5			127
Mixed10			142

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Table 7. Summary Statistics of Ratings (Study 1)

	Unit	Mean	Standard deviation	Minimum	Maximum
Study 1a					
Overall rating (SD)	1–9	6.23	1.27	2	9
Overall rating (MD)	1–9	6.75	0.93	4	9
Mean dimensional rating (MD)	1–9	7.17	0.70	5.25	9
Food	1–9	8.62	0.65	6	9
Service	1–9	8.67	0.66	6	9
Value	1–9	7.61	1.36	4	9
Atmosphere	1–9	3.79	1.77	1	9
Study 1b – Order Effects					
Overall rating (SD)	1–9	6.33	1.16	1	9
Overall rating (MD-first)	1–9	6.60	1.22	3	9
Overall rating (MD-last)	1–9	6.52	1.25	3	9
Overall rating (MD-prime)	1–9	6.51	1.04	2	9
Mean dimensional rating (MD-first)	1–9	6.85	0.83	1.5	9
Food	1–9	8.49	0.97	1	9
Service	1–9	8.67	0.80	3	9
Value	1–9	7.27	1.64	1	9
Atmosphere	1–9	2.98	1.65	1	9
Mean dimensional rating (MD-last)	1–9	7.08	0.90	3.5	9
Food	1–9	8.56	0.86	4	9
Service	1–9	8.66	0.76	3	9
Value	1–9	7.50	1.59	2	9
Atmosphere	1–9	3.61	1.90	1	9
Study 1c – Number of Dimensions					
Overall rating (SD)	1–9	6.44	1.39	3	9
Overall rating (Top5)	1–9	6.84	1.30	3	9
Overall rating (Mixed10)	1–9	6.58	0.94	5	9
Mean dimensional rating (Top5)	1–9	7.13	0.90	4	9
Food	1–9	8.60	0.83	3	9
Service	1–9	8.65	0.83	3	9
Value	1–9	7.65	1.55	1	9
Cleanliness	1–9	3.76	1.95	1	9
Price	1–9	6.98	1.79	2	9
Mean dimensional rating (Mixed10)	1–9	6.64	0.74	4.9	8.6
Food	1–9	8.55	0.69	6	9
Service	1–9	8.76	0.52	6	9
Value	1–9	7.46	1.35	3	9
Cleanliness	1–9	3.26	1.62	1	9
Price	1–9	6.84	1.69	2	9
Noise level	1–9	3.56	2.01	1	9
Waiters	1–9	8.32	1.04	2	9
Years in business	1–9	5.75	1.75	1	9
Wait time	1–9	7.43	1.69	1	9
Opening hours	1–9	6.48	1.63	1	9

suggesting that overall ratings in the MD condition deviate upward from those in the SD condition if mean dimensional ratings are *higher* than SD overall ratings. (See Table 4 for detailed results.)[18]

5.3. Study 1b: Order Effects

The purpose of Study 1b was to gain further insights into people's rating process. In particular, this study served two distinct purposes. First, we sought to examine order effects between ratings to test Hypothesis 2 (eliciting dimensional ratings first versus last). Following robust evidence from the literature on order effects, we expect no effect if the dimensional ratings are assessed *after* the overall rating—as is done by some websites—because the dimensions have not been elicited before forming the overall rating. Thus, we added another condition, in which we asked participants to provide their overall rating *before* the dimensional ratings.

Second, we sought to explore the effects of asking the participants only to consider the dimensions (rather than asking participants to rate them). To explore whether overall ratings can be influenced by merely highlighting certain dimensions (i.e., making dimensions more accessible without making participants elaborate on them), we added another condition: MD-prime. In this condition, we only presented the dimensions—without providing a scale for rating

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

them—and then asked the participants to provide an overall rating.[19]

5.3.1. Procedure and Measures. We asked participants in the MD-first/MD-last conditions to rate the individual dimensions before/after providing an overall rating. In the MD-prime condition, we first provided the following instruction: "Please take a moment to reflect on your visit to TASTY RESTAURANT, considering the following aspects: food, service, value, atmosphere"; following this, we asked the participants to provide an overall rating. Thus, we assigned participants to one of four conditions: (1) an overall-rating-only condition (SD condition, which served as the baseline), (2) dimensional ratings with subsequent overall rating (MD-first condition), (3) overall rating with subsequent dimensional ratings (MD-last condition), and (4) presentation of dimensions with subsequent overall rating but without asking participants to rate the individual dimensions (MD-prime condition). We adapted the model specification presented in Equation (1) by including the new conditions (β_3 = *MD-last*$_{overall}$, β_4 = *MD-last*$_{dimensional}$, *and* β_5 = *MD-prime*$_{overall}$).

In Study 1a, we elicited the dimensional ratings in the same order as TripAdvisor did, which is a potential limitation because participants might have been influenced primarily by the first or the last dimension (primacy/recency effect). To address this limitation, we randomized the order of dimensions.

5.3.2. Results. The purpose of Study 1b was to test Hypothesis 2 and to explore the effects of merely asking the participants to consider the dimensions (rather than asking to rate them). We first investigated whether the mean dimensional ratings in the MD conditions differed from the overall ratings in the SD condition. In line with Study 1a, we found that the mean dimensional ratings in the MD-first condition were significantly higher than the overall ratings in the SD condition (6.85 stars versus 6.33 stars, β_2 = .52, p < .001). Further, we found that the mean dimensional ratings in the MD-last condition were significantly higher than the overall ratings in the SD condition (7.08 stars versus 6.33 stars, β_4 = .75, p < .001). We then proceeded with examining the difference between overall ratings in the MD and SD conditions. As expected, we found that overall ratings in the MD-first condition were significantly higher than those in the SD condition (6.60 stars versus 6.33 stars, β_1 = .27, p = .011), lending further support to Hypothesis 1; interestingly, however, we found a marginally significant difference between the overall ratings in the MD-last condition and those in the SD condition (6.52 stars versus 6.33 stars, β_3 = .19, p = .068; see Table 7 for summary statistics). Finally, we found a marginally significant difference between overall ratings in the SD condition and MD-prime condition (6.51 stars versus 6.33 stars, β_5 = .18, p = .087). These results cast doubt on Hypothesis 2, which stated that dimensional ratings should not influence overall ratings if dimensional ratings are elicited after asking for overall ratings. Further, the results show that merely reminding people of several dimensions could have the potential to influence overall ratings.

5.4. Study 1c: Effect of Number of Dimensions

The purpose of Study 1c was to obtain further evidence for the accessibility–diagnosticity mechanism by testing Hypothesis 3, in which we proposed that the dimensional rating bias should decrease as the number of dimensions increases. Thus, we examined the difference between (1) a condition in which we elicited ratings on only five dimensions and (2) a condition in which we elicited ratings on 10 dimensions.[20]

5.4.1. Procedure and Measures. In a pretest (N = 100), we first told participants that service, food, value, and atmosphere were commonly regarded as important attributes when rating restaurants. We then asked the participants to list five *other* attributes they considered important during a restaurant visit. Next, we asked each participant to rank these nine attributes (service, food, value, and atmosphere and the five attributes provided by the participant) according to their importance. Based on these responses, we obtained a total of 105 attributes and a listing of the 10 most frequently listed attributes for each rank.

Using the results of the pretest, we created three conditions: (1) an overall-rating-only condition (SD condition), (2) a Top5 condition containing the five dimensions that were most frequently ranked as the most important (i.e., food, service, value, cleanliness, price), and (3) a Mixed10 condition containing the Top5 dimensions and five dimensions from among those being ranked lower by the participants (noise level, waiters, years in business, wait time, opening hours). We created five dummy variables to compare the rating scores: *SD*$_{overall}$ (overall rating from the SD condition, which served as the baseline), *Top5*$_{overall}$/*Mixed10*$_{overall}$ (overall rating from the Top5/Mixed10 condition), *Top5*$_{dimensional}$/*Mixed10*$_{dimensional}$ (mean dimensional rating from the Top5/Mixed10 condition). As in Studies 1a and b, we obtained one rating score for participants in the SD condition (β_0 = *SD*$_{overall}$) and two rating scores for participants in the MD conditions. We adapted the model specification

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

presented in Equation (1) using the conditions $\beta_1 = Top5_{overall}$, $\beta_2 = Top5_{dimensional}$, $\beta_3 = Mixed10_{overall}$, *and* $\beta_4 = Mixed10_{dimensional}$.

5.4.2. Results. We first investigated whether the mean dimensional ratings in the MD conditions differed from the overall ratings in the SD condition. We found that the mean dimensional rating in the Top5 condition was significantly higher than the overall rating in the SD condition (7.13 stars versus 6.44 stars, $\beta_2 = .69$, $p < .001$). In contrast, in the Mixed10 condition, we did not find a significant difference (6.64 stars versus 6.44 stars, $\beta_4 = .20$, $p = .132$), providing preliminary evidence that the effect of selectively highlighting only a few dimensions had diminished. We then proceeded with examining the difference between overall ratings in the MD and SD conditions. In the Top5 condition, the overall rating differed significantly from the overall rating elicited in the SD condition (6.84 stars versus 6.44 stars, $\beta_1 = .40$, $p = .004$). In the Mixed10 condition, the overall rating did not differ significantly from that of the SD condition (6.58 stars versus 6.44 stars, $\beta_3 = .14$, $p = .305$; see Table 7 for summary statistics). Together, the results support Hypothesis 3, suggesting that the overall rating in the Top5 condition was influenced by the selectively highlighted dimensions and that the effect diminishes as more dimensions are highlighted; this provides further support for our accessibility–diagnosticity argument.

5.5. Experiments on Restaurant Ratings: Discussion

The purpose of the restaurant experiments was to test our hypotheses in a controlled setting, that is, (1) whether eliciting dimensional ratings prior to an overall rating influences a subsequent overall rating (Hypothesis 1, Study 1a), (2) whether this effect disappears when dimensional ratings are assessed after overall ratings (Hypothesis 2, Study 1b), and (3) whether the dimensional rating bias decreases when the number of dimensions assessed prior to an overall rating increases (Hypothesis 3, Study 1c). In addition, we explored whether merely asking people to consider dimensions—without asking them to rate those dimensions—influences overall ratings (Study 1b).

The results of Study 1a support Hypothesis 1. In line with our accessibility–diagnosticity explanation, asking for ratings of selected dimensions led overall ratings between the SD and MD conditions to differ such that participants in the MD condition provided higher overall ratings than participants in the SD condition. Moreover, the direction of the difference suggests that these ratings were systematically influenced by the dimensions—operationalized using the mean of those dimensions.

In Study 1a, we asked for an overall rating *after* asking for dimensional ratings, whereas some websites ask for an overall rating *before* asking for dimensional ratings. We addressed these issues in Study 1b, which tested Hypothesis 2. To our surprise, the results of Study 1b—that overall ratings in the MD-last condition were higher (albeit marginally significantly) than SD ratings—suggest that dimensional ratings may have the potential to influence overall ratings even if we ask for dimensional ratings *after* asking for overall ratings, at least casting doubt on Hypothesis 2 and suggesting that participants may have adjusted their overall rating after providing the dimensional ratings. A similar effect was observed by Schwarz and Hippler (1995), who found that later questions affect earlier ones in situations in which respondents could go back and forth between questions. Unsurprisingly, this effect was smaller than the effect of dimensional ratings on subsequent overall ratings because it is likely that only some participants returned to the summary question to adjust their rating. Interestingly, the mean rating of the negatively framed dimension *atmosphere* was higher in the MD-last condition than in the MD-first condition (whereas the other dimensions remained fairly equal), suggesting a reciprocal influence in that the overall rating from the MD-last condition may also have influenced the rating of the specific dimensions; in other words, participants may have adjusted their rating of *atmosphere* upward so as to be aligned with their overall ratings.[21] In addition, we demonstrated that merely reminding people of several dimensions (without asking them to rate these dimensions) also has the potential to influence overall ratings. Together, the results of Study 1b provide further evidence that accessibility drives the provision of overall ratings.

Finally, we hypothesized that the dimensional rating bias would decrease as the number of dimensional ratings assessed increases such that the difference between MD overall ratings and SD ratings decreases (Hypothesis 3). Comparing the effects of asking participants to rate 5 versus 10 dimensions, the results of Study 1c show that the effect of dimensional ratings on overall ratings all but disappears if many dimensions are highlighted, supporting our accessibility–diagnosticity argument.

These results are in contrast to the findings of Chen et al. (2018), who found no significant priming effect of MD ratings on overall ratings in their experiment 2. A possible explanation is that their participants had to rate *experiences of others* (as operationalized by textual reviews); such mediated (secondhand) experiences are likely to differ from one’s own lived (firsthand) experiences, which “tend to be stronger than ‘secondhand’ experiences in determining consumers’ notion

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

of reality" (Whelan and Wohlfeil 2006, p. 316). Whereas in both our study and that of Chen et al. (2018), participants did not have a fully lived experience, rating a restaurant on the basis of others' reviews (i.e., mediated or secondhand experiences) is likely to have an even higher experience gap than rating a fictitious restaurant visit, especially because the reviews presented in the study by Chen et al. (2018) already included evaluations by others. Nevertheless, the limitation that participants in Study 1 had to rate a fictitious scenario remains. In other words, this study did not expose participants to a lived experience, which may have influenced their rating behavior, so we conducted further experiments (Studies 2 and 3) that more closely approximated an actual rating situation in order to replicate Study 1a, further test Hypothesis 1, and explore the effects on varying the dimensions elicited.[22]

6. Study 2: Experiments on Movie Ratings

6.1. Overview

In the previous experiments, we used a fictitious scenario, which represented to some extent a mediated experience; thus, we conducted Studies 2a–c using lived experiences—short films—as rating objects to replicate and extend our findings. Further, in Study 1, MD overall ratings were always higher than SD overall ratings. Likewise, Chen et al. (2018), using information-transfer theory, predicted and found that MD overall ratings were always higher. Following Feldman and Lynch's (1988) accessibility–diagnosticity framework, however, MD overall ratings could also be lower than SD overall ratings if the dimensional ratings are related to comparatively more negative experiences. To test this aspect of Hypothesis 1, we conducted Study 2b using a short film for which we expected the dimensions to be rated lower than the overall ratings in the SD system. Finally, we conducted Study 2c to find further evidence for the proposed mechanism—dimensional ratings influence overall ratings—by examining the effects of asking for dimensions that we expected to score very high or low in valence (i.e., essentially, we manipulated the mean of the dimensional ratings). As in Study 1, we first describe the elements that were common across the experiments; we then present the specific details of each study before providing a general discussion of the findings.

6.1.1. Procedure and Materials. To ensure that participants did not engage in hypothesis guessing, we told them the study was about visual awareness. We asked them to watch a short film and told them they would be asked several questions about it. Participants watched either *Canhead*, a seven-minute, stop-motion film (Study 2a), or *Xiao Xiao*, a two-minute animated martial arts film (Studies 2b and c).[23] Sessions in Study 2a (*Canhead*) lasted an average of 15 minutes, whereas sessions in Study 2b and c (*Xiao Xiao*) lasted an average of eight minutes. Dimensional ratings were based on the dimensions used by Yahoo! Movies: *visual effects*, *animation quality*, *stream quality*, and *story*.[24] Based on a pretest,[25] we expected the short film in Study 2b (*Xiao Xiao*) to be rated low on the dimensions *visual effects*, *animation quality*, and *story*. After the experiment, we collected demographic data. We conducted the entire study using the Qualtrics survey platform.

6.1.2. Measures and Model Specification. The variables and measures were equivalent to those described in Study 1 (i.e., three rating scores: $SD_{overall}$, $MD_{overall}$, and $MD_{dimensional}$ as the mean rating of the dimensions *visual effects*, *animation quality*, *stream quality*, and *story*). We created dummy variables to compare our rating scores. We used the same model specification as that presented in Equation (1) for Studies 2a and b and used *t*-tests to compare the overall ratings of the two conditions in Study 2c.[26]

6.1.3. Participants. As in Study 1, we used G*Power to determine the sample size (Faul et al. 2009). We used Prolific to recruit participants who met the following criteria: reside in English-speaking countries (Australia, Canada, England, New Zealand, and the United States), be at least 18 years old, have an approval rating of at least 90% for previously completed studies, and have not participated in any of our other studies. We paid all participants £1 (about US$1.25) for a 10-minute task, a payment equivalent to an hourly wage of £6 (US$7.50). We randomly assigned the participants to conditions using Qualtrics' Randomizer. We removed any incomplete cases from the analyses.

Table 8. Study 2: Participant Demographics

	Study 2a	Study 2b	Study 2c (pilot)	Study 2c (main)
Participants (recruited)	200	200	500	1,000
Participants (analyzed)	192	198	408	847
Average age	31.4	32.5	35.7	31.9
Women, %	42.7	45	48.3	49

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Table 8 provides details about the participant demographics. Table 9 provides details about the number of participants in the different conditions.

6.2. Study 2a: Positive Influence of MD Ratings

The purpose of Study 2a was to replicate Study 1a in a more realistic setting. We used the materials, procedures, and model specifications outlined in the preceding overview.

As in Study 1, we first investigated whether the mean dimensional ratings in the MD condition differed from the overall ratings in the SD condition; we found that the mean dimensional ratings in the MD condition were significantly higher than the overall ratings in the SD condition (6.68 stars versus 5.87 stars, $\beta_2 = .81$, $p = .002$; see Table 10 for summary statistics). We then proceeded with testing our hypothesis and found that overall ratings in the MD condition were higher than those in the SD condition; although this was the direction we expected, the difference was not significant (6.02 stars versus 5.87 stars, $\beta_1 = .15$, $p = .558$), so we did not find further support for our hypothesis. (See Table 4 for detailed results.)

6.3. Study 2b: Negative Influence of MD Ratings

The purpose of Study 2b was to further test Hypothesis 1 and to examine whether MD overall ratings could also be lower than SD overall ratings if the dimensional ratings were related to comparatively more negative experiences. We used the materials, procedures, and model specifications outlined in the preceding overview.

As in Study 2a, we first investigated whether the mean dimensional ratings in the MD condition differed from the overall ratings in the SD condition. We found that the mean dimensional ratings in the MD condition were significantly lower than the overall ratings in the SD condition (5.05 stars versus 5.72 stars, $\beta_2 = -.67$, $p = .027$; see Table 10 for summary statistics). We then tested our hypothesis and found that overall ratings in the MD condition were also significantly lower than those in the SD condition (4.69 stars versus 5.72 stars, $\beta_1 = -1.03$, $p < .001$). Thus, the results of Study 2b support our hypothesis that overall ratings from an MD system will be lower than overall ratings from an SD system if the mean of the dimensional ratings is lower than the mean of the SD overall ratings.

Table 9. Study 2: Participants per Condition

	Study 2a	Study 2b	Study 2c
Participants (analyzed)	192	198	847
SD	91	92	
MD	101	106	
MD-NoStory			425
MD-NoStream			422

6.4. Study 2c: Effect of Manipulating Dimensions

The purpose of Study 2c was to find further evidence for the proposed mechanism by examining the effects of varying the content of the dimensions (i.e., essentially, manipulating the mean of the dimensional ratings by asking for dimensions that we expected to differ in valence). Based on Study 2b, we created subsets of dimensions that included or omitted dimensions that would be rated relatively high or low. According to our theoretical argument, we expected the overall rating to be higher if we excluded a dimension that was rated relatively low (e.g., *story*); conversely, we expected the overall rating to be lower if we excluded a dimension that was rated relatively high (e.g., *stream quality*).[27]

6.4.1. Procedure, Materials, and Conditions. We used the same materials as described in Study 2b and adjusted the conditions to vary the content of the dimensions. Whereas in Study 2b we used *one* dimensional-rating condition with four dimensions—visual effects, animation quality, stream quality, and story—in Study 2c, we used a total of *four* dimensional-rating conditions, each of which excluded *one* of the mentioned dimensions (see Table 11). According to the proposed mechanism, the overall rating should be higher if we exclude a dimension rated extremely low or, conversely, should be lower if we exclude a dimension that was rated extremely high.

6.4.2. Pilot Study. The results of Study 2 suggested that *story* and *stream quality* were the lowest and highest rated dimensions, respectively. We conducted a pilot study to confirm that the conditions omitting these dimensions were indeed the appropriate conditions for use in our main experiment and to determine the sample size needed for the main experiment. In particular, we included a total of four dimensional-rating conditions, each of which excluded one of the mentioned dimensions, and determined in which of the four conditions the overall ratings were most extreme; subsequently, we used the effect size to calculate the sample size for our main experiment.

The results of this pilot study show that the overall rating was highest in the condition without story (5.57 stars) and that the story dimension was consistently rated relatively low in the other conditions (4.23, 4.44, and 3.45); conversely, the overall rating was lowest in the condition without stream quality (5.00 stars), and the stream quality dimension was rated relatively high in the other conditions (7.41, 7.53, and 7.19; see Table 12). The results of this pilot study are consistent with the results of Study 2b and indicate that extreme

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Table 10. Study 2: Summary Statistics of Ratings

	Unit	Mean	Standard deviation	Minimum	Maximum
Study 2a (*Canhead*)					
Overall rating (SD)	1–9	5.87	1.95	1	9
Overall rating (MD)	1–9	6.02	1.91	1	9
Mean dimensional rating (MD)	1–9	6.68	1.36	3.25	9
Visual effects	1–9	6.50	1.75	2	9
Animation quality	1–9	6.84	1.91	1	9
Stream quality	1–9	7.59	1.64	2	9
Story	1–9	5.78	2.04	1	9
Study 2b (*Xiao Xiao*—all dimensions)					
Overall rating (SD)	1–9	5.72	2.21	1	9
Overall rating (MD)	1–9	4.69	2.24	1	9
Mean dimensional rating (MD)	1–9	5.05	1.80	1	9
Visual effects	1–9	4.37	2.35	1	9
Animation quality	1–9	5.51	2.45	1	9
Stream quality	1–9	7.14	2.13	1	9
Story	1–9	3.18	2.11	1	9
Study 2c (*Xiao Xiao*—w/o story)					
Overall rating	1–9	5.70	2.16	1	9
Mean dimensional rating	1–9	6.24	1.74	1	9
Visual effects	1–9	5.25	2.29	1	9
Animation quality	1–9	6.09	2.21	1	9
Stream quality	1–9	7.37	1.92	1	9
Study 2c (*Xiao Xiao*—w/o stream)					
Overall rating	1–9	4.95	2.10	1	9
Mean dimensional rating	1–9	4.79	1.95	1	9
Visual effects	1–9	4.99	2.32	1	9
Animation quality	1–9	5.74	2.37	1	9
Story	1–9	3.64	2.28	1	9

dimensions influence overall ratings. We used G*Power (Faul et al. 2009) to calculate the sample size required to detect a significant difference between the overall ratings in the conditions without story and without stream quality. Based on an effect size of $d = 0.24$, $\alpha = 0.05$, and Power = 0.95, we determined that, to observe a significant difference, we required a sample size of 439 participants per condition.

6.4.3. Main Experiment. We proceeded with only the two conditions of interest for the main experiment—the condition without story (MD-NoStory) and the condition without stream quality (MD-NoStream) and compared the overall ratings using an independent t-test.

6.4.4. Results and Discussion. The results of the main experiment show that the overall rating in the MD-NoStory condition was significantly higher than the overall rating in the MD-NoStream condition (5.70 stars versus 4.95 stars, $t(845) = 5.145$, $p < .001$; see Table 10 for summary statistics). In other words, participants provided overall ratings that differed on average 0.75 stars between these conditions, indicating that the overall rating is highly dependent on the dimensions and can be manipulated by omitting certain dimensions.

6.5. Experiments on Movie Ratings: Discussion

Studies 2a–c were designed to demonstrate the effects of dimensional ratings on overall ratings using lived

Table 11. Study 2c: Dimensions Included in Each Condition

	Condition				
	Without visuals	Without animation	Without stream	Without story	OR only
Overall rating	×	×	×	×	×
Visual effects		×	×	×	
Animation quality	×		×	×	
Stream quality	×	×		×	
Story	×	×	×		

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Table 12. Study 2c (Pilot Test): Average Ratings

	Condition				
	Without visuals	Without animation	Without stream	Without story	OR only
Overall rating	5.28 (2.20)	5.46 (2.61)	5.00 (2.20)	5.57 (2.47)	5.39 (2.33)
Visual effects		5.50 (2.78)	5.19 (2.36)	5.21 (2.65)	
Animation quality	5.38 (2.45)		5.62 (2.38)	5.79 (2.54)	
Stream quality	7.41 (1.64)	7.53 (1.86)		7.19 (2.15)	
Story	4.23 (2.47)	4.44 (2.83)	3.45 (2.40)		

Note. Standard deviation in parentheses.

experiences—short films—as rating objects so as to replicate and extend our findings. Whereas we designed Study 2a as a replication, we conducted Study 2b to further test Hypothesis 1 by using a short film for which we expected the dimensions to be rated lower than the overall ratings in the SD system. Further, we conducted Study 2c to find further evidence for the proposed mechanism by exploring the effects of asking for certain dimensions that would be expected to differ in valence (i.e., essentially, manipulating the mean of the dimensional ratings).

Together, the results of Studies 2a–c provide further evidence that dimensional ratings influence subsequent overall ratings. Unexpectedly, the difference between overall ratings in Study 2a was not significant (per definition, however, nonsignificant results do not represent evidence *against* a hypothesis), yet the direction of the overall rating in the MD condition tended toward the mean of the dimensional ratings. One potential explanation for this finding could be that, whereas we assumed equal weighting of dimensions in our analysis, participants did not use equal weighting when arriving at their overall rating. As Table 10 shows, the ratings of the different dimensions were very divergent with *stream quality* rated very high and *story* rated comparatively low (5.78, which is close to the overall rating of 5.87 in the SD condition), which suggests that the dimension *story* might have been weighted more heavily by the participants (see also the following discussion of Study 2b).

The results of Study 2b show that MD ratings can influence overall ratings in both directions: if the mean dimensional rating in an MD system is higher than the overall rating in an SD system, overall ratings in an MD system deviate in that direction (Studies 1a and b), whereas if the mean dimensional rating in an MD system is lower than the overall ratings in an SD system, overall ratings in the MD system deviate in this direction (Study 2b), challenging the claim of Chen et al. (2018) that dimensional ratings should *always* lead to higher overall ratings. Interestingly, the overall rating in the MD condition was even more extreme than the mean dimensional rating. There are two possible explanations for this finding. First, the participants may have weighted the dimensions differently. Whereas we calculated the mean rating by using equal weights for each dimension, real (unobservable) weightings are likely not equal across dimensions. If a dimension with a rating that is more extreme than the mean (such as *story*) has a high weight, the overall rating overshoots the mean rating. Second, the accessibility of experiences that were not assessed during the dimensional ratings might also depend on these ratings. Valence-consistent memories might be more accessible than valence-inconsistent memories, leading to an overshoot in the overall rating.

Finally, the results of Study 2c demonstrate that overall ratings are susceptible to changes to the dimensions included in MD rating systems. Our results show that overall ratings differ significantly if certain (extreme) dimensions are highlighted or omitted—that is, if a few dimensions are made more accessible—and this finding, therefore, supports our theoretical explanation of accessibility. In other words, we were able to actively manipulate the overall rating by including or omitting certain dimensions, which suggests that, in MD systems, the content of the dimensions can bias overall ratings—a finding that has tremendous practical implications for the design of rating systems.

Although participants in Study 2 were exposed to real-world experiences, the experience (i.e., watching short films) was rather short, and the participants' level of involvement was likely to be low. In addition, we have, thus far, shown the effect only using carefully controlled scenarios and stimuli. We designed Study 3 as a field experiment to address these limitations.

7. Study 3: Field Experiment on University Ratings

7.1. Overview

The results of Studies 1 and 2 provide evidence that, when controlled stimuli are used, dimensional ratings may influence overall ratings. In Study 3, we sought to test Hypothesis 1 in a real-world setting by asking students to rate their own university, thus addressing ecological validity. Further, the "experiences"

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

to be rated in Studies 1 and 2 were of relatively short duration, and participants' level of involvement with the rating targets was likely to be low. In contrast, a student's involvement with the university is likely to be higher; further, the experience is of longer duration, allowing the participants to form an evaluation with higher certainty. In line with our previous findings, we expected the overall university ratings to differ between SD and MD rating systems; in particular, we expected that in the MD condition, the overall rating would be biased toward the mean of the dimensional ratings (i.e., higher (lower) than overall ratings from an SD system if the mean of the dimensional ratings is higher (lower) than the mean of the SD overall ratings).

7.2. Participants

We recruited 150 students from a public university in Europe. The participants completed the study during lecture time using their own devices; we did not provide any compensation for participation. Following Studies 1 and 2, we used Qualtrics' Randomizer to assign participants to the conditions. We removed from the analysis eight participants with incomplete data, resulting in a final sample size of 142 participants, 77 of whom were in the MD condition. All participants were over 18 years old; the mean age was 24.0 years, and 36% were women.

7.3. Procedure and Measures

We created a website containing the experimental stimuli using Qualtrics.com and disseminated the link among participants by sharing it at lectures for undergraduate business students. After presenting the instructions, the website displayed a picture of the university, followed by a distractor movie.[28] Then, participants in the MD condition had to provide dimensional ratings of their university, followed by an overall rating, which all participants had to complete (in both the MD and SD conditions). After the experiment, we gathered demographic data. The sessions lasted 10 minutes on average (the materials can be found in e-companion section EC.2).

We asked the participants in the MD condition to rate their university on the dimensions *teaching*, *value for money*, *internationality*, *student service*, *campus*, and *security* (the dimensions used by the university rating site www.study-advisor.org),[29] followed by an overall rating. Participants in the SD condition had to provide an overall rating only. We used the same model specification, described in Equation (1), using the following conditions: $SD_{overall}$ (β_0, which served as the baseline), $MD_{overall}$ (β_1), and $MD_{dimensional}$ (β_2).

7.4. Results and Discussion

In line with Studies 1 and 2, we first tested whether the mean dimensional ratings differed from the overall ratings in the SD condition and found that they were significantly higher, albeit only marginally so (6.65 stars versus 6.15 stars, $\beta_2 = .50$, $p = .055$; see Table 13 for summary statistics). We then tested our hypothesis—we expected the overall rating from the MD condition to be higher as well—and found that overall ratings in the MD condition were significantly higher than those in the SD condition (6.87 stars versus 6.15 stars, $\beta_1 = .72$, $p = .006$), further supporting Hypothesis 1. (See Table 4 for detailed results.) As in Study 2b, the overall rating in the MD condition was even more extreme than the mean dimensional rating; this might be attributable to differential weighting of the dimensions (e.g., the study fees were relatively low, so *value for money* might have been weighted more heavily by the respondents). Addressing ecological validity, the results from this real-world study provide further evidence that MD ratings systematically influence overall ratings in MD systems.

8. General Discussion

8.1. Summary

Electronic word-of-mouth communication, especially in the form of online ratings, can influence consumer decision-making. However, prior studies have shown that ratings can be biased by self-selection and social influence. In this study, we examined the dimensional-rating bias, that is, the influence of eliciting MD ratings on overall ratings. Our study was

Table 13. Study 3: Summary Statistics of Ratings

	Unit	Mean	Standard deviation	Minimum	Maximum
Overall rating (SD)	1–9	6.15	1.83	1	9
Overall rating (MD)	1–9	6.87	1.29	3	9
Mean dimensional rating (MD)	1–9	6.65	1.27	2.67	9
Teaching	1–9	6.39	1.62	1	9
Value for money	1–9	6.97	1.96	1	9
Campus	1–9	5.56	2.30	1	9
Security	1–9	7.87	1.63	2	9
Internationality	1–9	6.66	1.99	2	9
Student service	1–9	6.44	1.92	2	9

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

guided by the following research question: how do MD ratings influence overall ratings?

Using seven experiments with a total of 17 conditions (see Table 3 for an overview) as well as an additional study presented in e-companion section EC.1, we demonstrate that people's overall ratings are influenced systematically by dimensional ratings. In particular, overall ratings tended toward the mean of the dimensional ratings, suggesting that dimensional ratings systematically influence overall ratings. In Studies 1a, 1b, and 3, the means of the dimensional ratings in the MD condition were higher than the overall ratings in the SD condition as were the overall ratings in the MD condition (note that we did not find a significant difference in Study 2a). In contrast, in Study 2b, the mean of the dimensional ratings was lower than the overall rating in the SD condition as was the overall rating in the MD condition.

Together, our series of experiments not only examined boundary conditions but also offered further insights into the rating process. First, MD ratings can influence overall ratings not only upward but also downward. Second, contrary to our expectation, MD ratings might have the potential to influence overall ratings even if they are elicited *after* the overall rating, suggesting that participants may have adjusted their overall rating after providing the dimensional ratings; this influence might be reciprocal in that the overall rating may also have influenced the rating of individual dimensions. Third, MD ratings have the potential to influence overall ratings even if raters are *only reminded* of dimensions without having to evaluate them. Finally, MD ratings can systematically bias overall ratings, *depending on the dimensions* being included.

Refer back to Table 4 for a comparison of the results of our experiments. In all of our experiments, the overall ratings in the MD condition tended toward the mean of the dimensional ratings, providing suggestive evidence of our proposed mechanism. To provide further evidence for the proposed mechanism, we conducted an additional post hoc analysis. Following our proposed explanation, we expect that the overall rating in an MD system is influenced by the dimensional ratings, such that the overall rating tends toward the dimensional rating's mean; consequently, the effect should be stronger for respondents who provided more extreme dimensional ratings. To test this presumption, we reanalyzed the data from Studies 1, 2a, 2b, and 3. Using the data of each study, we created two subsets, containing (1) responses with extreme dimensional ratings, data that included only the observations in the bottom 25% and top 25% of mean dimensional ratings, and (2) responses with moderate dimensional ratings, the middle 50%. We then calculated the correlations between mean dimensional ratings and overall ratings (both from the MD conditions). As expected, the correlations were significantly stronger in the extreme observations (see the appendix for the results of this analysis). Together with our main results, these results suggest that dimensional ratings make these dimensions more accessible, thereby influencing overall evaluations.

8.2. Scholarly Implications, Limitations, and Future Research

Our results contribute to the understanding of online ratings. Prior research on rating biases has focused primarily on the self-selection bias or social influence (Hu et al. 2009, Muchnik et al. 2013), devoting little attention to effects that may arise from the rating method itself.[30] We fill a gap in the rating literature by showing that the design of the rating platform—specifically, the platform provider's decision to use an MD or an SD rating system—influences overall ratings. Prior studies have used information-transfer theory and expectation-confirmation theory to explain that MD systems always result in overall ratings that are higher than the ratings from SD systems (Chen et al. 2018). In contrast to the findings of Chen et al. (2018) (in particular, in contrast to the nonsignificant results of their priming experiment), we demonstrated—in a number of studies conducted across different settings—that the design of the rating systems itself can influence ratings: by making certain dimensions more accessible, MD systems can lead to either upward or downward deviations in overall ratings, depending on the means of the dimensional ratings. Thus, we contribute to the research on another source of bias in online ratings: the *dimensional-rating bias*. Further, whereas prior research has consistently shown strong order effects from preceding to subsequent questions, our results show that the influence may go in the reverse direction as well; a similar effect was demonstrated in a study by Schwarz and Hippler (1995), who argued that respondents may have gone back to adjust their answers. Likewise, we observed that, in online rating contexts, not only do dimensional ratings influence overall ratings, but overall ratings have the potential to influence subsequent dimensional ratings. These findings show that online rating behavior is complex, and we call on researchers to further explore the reciprocal effects of overall ratings on dimensional ratings.

Our experiments are not without limitations. First, our experiments provide ample support for the underlying theoretical mechanism, but provide only suggestive evidence and do not test the mechanism per se. Although our additional post hoc analysis provides some evidence for the suggested

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

mechanism, future research could devise ways to test the theoretical mechanisms explicitly (such as by introducing other manipulations or using neurophysiological tools to examine the underlying cognitive processes).

Second, although we used a range of rating objects (restaurants, short films, universities), they were primarily of an experiential nature, and the ratings were concerned with intangible experiences. We did not examine whether our findings could be applied to tangible products. Likewise, future research could focus on utilitarian (as opposed to hedonic) products. Future research could also compare the effects in the context of search goods (in which product features or qualities can be evaluated before consumption) versus experience goods (in which certain qualities are difficult to evaluate prior to consumption) (Darby and Karni 1973).

Third, although we demonstrated that merely asking people to rate particular dimensions influences overall ratings, we did not seek to determine *how* participants may weight individual dimensions. In other words, different dimensions may be weighted differently (see Studies 2b and 3). Future research could explore how different dimensions are weighted and to what degree weightings may differ across different categories of products or services in order to determine the optimal set of dimensions for a given category.

Fourth, our studies did not address issues related to participants' decisions regarding whether to provide ratings.[31] However, we believe that the set of experiments was suitable for demonstrating the dimensional rating bias and explains the behaviors that emerge after a person has decided to provide a rating.

Finally, although we demonstrated that the effects hold in a variety of contexts, we conducted the majority of our studies (with the exception of Study 3) using scenarios that were controlled to varying degrees, and the experimental nature of the studies may have influenced participants' responses. Although we believe that the set of experiments presented here served as a strong test of our hypotheses, future research should replicate our studies in the setting of "live" commercial rating platforms.

Our studies shed further light on people's behaviors in online rating systems, yet a single paper can never address all aspects of a phenomenon. We call on researchers to build on our findings and examine other important avenues of analysis. For example, although it might be impossible to determine whether it is the SD or MD ratings that are "biased," researchers could attempt to discover and test measures that help resolve this issue; for example, potential objective measures could include ratings given directly after the experience versus delayed ratings, the stability of the ratings (i.e., test–retest reliability), or the predictive capacity of the ratings (in terms of repurchase behavior or in determining purchasing intentions of future customers).[32] By using dimensions used by major review sites, our studies (by design) primarily included dimensions that were likely to be high in diagnosticity; future research could use such objective measures to attempt to disentangle the role of accessibility versus diagnosticity of the individual dimensions in producing biased ratings. For example, do the observed effects result from the higher accessibility of the dimensions, or are they weighted more heavily because of higher diagnosticity? Relatedly, to what degree do reviewers assume that the dimensions included are diagnostic? Further, as suggested by Study 2b, summary ratings might influence dimensional ratings in MD systems (especially when the individual ratings are elicited following the summary ratings as was the case at TripAdvisor.com at the time of this writing); although it stands to reason that some respondents returned to the prior question to adjust their rating after providing dimensional ratings, future research could attempt to examine these reciprocal influences (e.g., using process data such as mouse-tracking data) in order to obtain an even deeper understanding of online ratings. Clearly, fruitful avenues for future research abound.

8.3. Implications for Practice

Our findings have important implications for different stakeholder groups. For any rating system, we can identify three main groups of stakeholders: (1) the provider of the rating system, (2) the product/service providers, and (3) the consumers. These three stakeholder groups have different goals, and our findings have important implications for each of them. In the following sections, we briefly highlight the goals and discuss the practical implications for each of these groups.

8.3.1. Implications for Providers/Designers of Rating Systems. Rating systems help customers learn about the quality of products or services (Gao et al. 2015) so that they can make informed decisions. Arguably, providers of rating systems strive to be viewed as a trusted provider of (quality) reviews to attract customers and monetize the web traffic. Hence, these systems use various elements and information to minimize uncertainty (Ba and Pavlou 2002) and maximize trust (Resnick et al. 2000). For example, dimensional ratings minimize uncertainty by providing a more complete picture of the product or service (Chen et al. 2018), helping customers to conduct prepurchase evaluations based on the ratings that are most important in their own situations or contexts.

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

However, our research shows that designers of rating systems should carefully consider the consequences of using dimensional ratings. We show that dimensional ratings influence subsequent overall ratings (Studies 1a, 1b, and 3), that this effect can be negative as well (Study 2b), and that the dimensional ratings might influence overall ratings even if they are rated after the overall ratings (Study 1b). Further, our results demonstrate that the influence of the dimensional ratings on overall ratings is heavily dependent on the dimensions rated—and, conversely, the dimensions omitted (Study 2c; see also Studies 2a, 2b, and 3). Thus, the mutual influences between SD and MD ratings show that it is likely impossible to arrive at an evaluation that accurately reflects raters' level of satisfaction, and providers of MD systems need to evaluate (such as by using large-scale testing) the right mix of dimensions that best reflect the quality of the product or service.

If providers decide to use a dimensional rating system, they should also be aware of the trade-off between using dimensional ratings and the reviewers' effort. Dimensional ratings can help to provide a more complete picture and minimize uncertainty. In addition, the results of Study 1c demonstrate that, with an increasing number of dimensions, the effect of rating these dimensions on overall ratings diminishes. This implies that rating too few dimensions increases the likelihood that the overall ratings will be biased. However, MD ratings require considerably more cognitive effort during the rating process (and rating multiple dimensions takes more time). It stands to reason that this higher cognitive (and time) burden can discourage people from providing ratings. In SD systems, the effort needed is considerably lower, which is less likely to discourage people from providing ratings. Designers of rating systems should conduct A/B tests to determine in how far increasing the number of dimensions elicited influences the number of ratings provided (e.g., McCloskey 2015, Siroker and Koomen 2015).

Alternatively, providers of rating systems may choose to ask users to provide either SD ratings only (as in the case of Yelp) or dimensional ratings only (without asking for overall ratings). Asking users to provide only one type of rating would mitigate the potential reciprocal influence of ratings in systems that combine dimensional and overall ratings. When asking reviewers to provide only dimensional ratings (without an overall rating), the rating platforms could then aggregate these to calculate a summary score. In calculating this score, the rating platforms could even allow readers to assign weights to different dimensions to account for different preferences regarding the dimensions; alternatively, rating platforms could analyze user behaviors or preferences in order to dynamically calculate summary values based on inferred preferences. As noted by de Langhe et al. (2016), "consumers' trust in the average user rating as a cue for objective quality appears to be based on an 'illusion of validity'" (p. 817). Dynamically constructing summary scores to account for consumers' preferences can be a step toward alleviating this problem.

With their influence on consumer buying behavior, providers of rating systems are in a unique position to bring about societal change and increase consumer welfare. Because dimensional ratings influence the overall ratings in the direction of the criteria expressed in MD rating systems, buying behavior is eventually influenced by such dimensions. For instance, by adding rating dimensions on, for example, the ecological or societal footprint of a product or service, the overall rating of products and services that score higher on these dimensions will receive a higher overall rating and will, in turn, be chosen more often by buyers; at the same time, introducing such rating dimensions would encourage product/service providers to adjust their offerings in a way to score particularly high on such dimensions in order to also positively influence the overall rating of their product or service (see our implications for product/service providers).

Although it is likely impossible to arrive at perfectly accurate evaluations, we advise providers of rating systems to be very transparent about the origins of the scores presented in order to allow consumers to critically evaluate the scores, which can, in turn, increase consumers' trust in both the ratings and the platform.

8.3.2. Implications for Product/Service Providers.

Arguably, a primary goal of providers (or producers) of goods and services is to obtain (many) good ratings because more positive ratings can increase the demand for the product/service (Chevalier and Mayzlin 2006). Consequently, given the demonstrated effects of dimensional ratings on overall ratings (Studies 1, 2b, 2c, and 3), product/service providers should not only focus on what they believe matters to their customers (or what customers tell them they care about), but should also be aware of what dimensions are being included by the major rating systems. The results of Study 2c demonstrate with particular clarity that the effects of dimensional ratings on overall ratings depend heavily on the questions asked and/or omitted. Although different aspects of the experience matter for different consumers (reviewers), any dimension included by the rating system has the potential to influence overall ratings. Consequently, being aware of the dimensions rated—and excelling on those dimensions—is likely to result in higher overall ratings.

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

8.3.3. Implications for Consumers. Finally, our findings have important implications for consumers as well. Although consumers are likely interested in obtaining unbiased ratings that reflect the "truth," arriving at such truth may not be possible. Because overall ratings differ systematically between SD and MD rating platforms (Studies 1, 2b, and 3), consumers need to be aware that overall ratings differ between rating systems (MD versus SD) and are, thus, incomparable. Our findings also demonstrate that overall ratings from different MD systems are not necessarily comparable because the ratings are likely to be influenced by the dimensions highlighted in the different rating systems, which can either raise or lower an overall rating (Study 2c); moreover, because the mere mention of dimensions has the potential to influence ratings (Study 1b), even ratings between different SD systems may not be comparable. Therefore, consumers should compare ratings only from the same platforms.

Further, consumers' preferences are likely to differ based on general preferences or situational factors (such as visiting a restaurant with sports buddies as opposed to a romantic partner). Consequently, users of MD rating systems should evaluate the match between the dimensions listed and their own attribute preferences because the degree to which the dimensions reflect their preferences or value system reveals the degree to which the overall rating on the rating platform is relevant to their own decision. It is impossible to recommend the best system for consumers, but we advise consumers to (1) choose systems that are open and transparent about how they construct overall ratings, and (2) critically evaluate the different systems in order to identify the system that can provide ratings that are most helpful in selecting the most suitable product or service.

In sum, our findings have various implications for designers of rating platforms, product/service providers, and consumers. Further, it stands to reason that our findings apply to contexts beyond rating platforms. For example, the dimensional-rating bias may also affect online evaluation/feedback forms and other surveys. A company that elicits feedback about its products should take these potential biases into account when interpreting the results of such surveys.

9. Conclusion

The results of our experiments suggest that consumers' ratings on particular dimensions significantly influence their overall ratings. If dimensional ratings are high, the overall rating tends to be higher as well, but if dimensional ratings are low, the rating tends to be lower. In addition, this effect depends heavily on the dimensions rated. Together, these results indicate a dimensional-rating bias; they show that information-systems design can influence rating behavior and that small modifications to an online choice environment may have a significant effect on consumer behavior.

Acknowledgments

The authors thank Raphael Ender, Inga Hämmerle, and Luiza Hoxaj from the University of Liechtenstein for outstanding research assistance. They also thank participants of research seminars at Erasmus University Rotterdam, University of Cologne, University of Konstanz, University of Mannheim, University of Paderborn, and Queensland University of Technology for their valuable comments. They also gratefully acknowledge the support from the following sources: IESE Business School; Erasmus University Rotterdam; the University of Liechtenstein; and Digital Innovation Laboratory of the Department of Information Systems, City University of Hong Kong. First authorship was determined by coin flip. Christoph Schneider and Markus Weinmann contributed equally to this work: experimental design, data collection, and analysis as well as introduction, background, method, results, and discussion; Peter N.C. Mohr contributed to the background, theory and hypotheses, and the theoretical implications section; Jan vom Brocke contributed to the practical implications section. Finally, the authors thank the senior editor, associate editor, and the anonymous reviewers for their valuable comments on earlier versions of this manuscript.

Appendix. Further Analysis to Test Causal Mechanism

In a further analysis, we examined whether extreme average ratings (in either direction) have a stronger influence on the overall rating.[34] To test this, we reanalyzed all experiments separately (see Table A.1). For this analysis, we created two subsets. The first subset included the participants who had provided extreme average dimensional ratings—that is, it included only the observations in the bottom 25% and the top 25% (i.e., 0.25 and 0.75 quantile; Q1 and Q4). The second subset included the remaining data—that is, the average dimensional ratings ranging from 0.25 to 0.75 (Q2 and Q3). Next, we calculated the correlations between the overall rating and the mean dimensional rating, which we expected to be stronger in the extreme data set; this was indeed the case—the correlations were larger in all cases and significantly different from those in the moderate subset in all studies.

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Table A.1. Correlation Between Overall Ratings and Mean Dimensional Ratings (Both from MD Condition)

		Extreme subset Q1 and Q4	Moderate subset Q2 and Q3	Compare correlations
Study	Experiment	r	r	z
1a	Restaurant (same page)	0.67^{***}	0.08	3.18^{***}
1b	Restaurant (MD-last)	0.74^{***}	0.25^{*}	4.63^{***}
1b	Restaurant (MD-first)	0.54^{***}	0.33^{**}	1.73^{*}
1c	Restaurant (Top5)	0.75^{***}	0.13	4.43^{***}
1c	Restaurant (Mixed10)	0.66^{***}	0.27^{*}	2.99^{***}
2a	Movie (*Canhead*)	0.89^{***}	0.33^{*}	5.17^{***}
2b	Movie (*Xiao Xiao*)	0.94^{***}	0.49^{***}	5.83^{***}
3	University	0.82^{***}	0.39^{**}	2.93^{***}

Note. Q1, Q2, Q2, and Q4 are quartiles.
$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Endnotes

[1] Typically in the form of star ratings and/or numerical scores.

[2] We replicated the study conducted by Chen et al. (2018) by comparing restaurant ratings from the rating platforms Yelp—currently an SD rating platform—and TripAdvisor—currently an MD rating platform; like Chen et al. (2018), we demonstrated that—in a real-world setting—overall ratings differ between SD and MD platforms. Details of this replication study are provided in the e-companion to this paper (Section EC.1).

[3] Although the expected effect is similar to our proposed dimensional rating bias, the authors did not elaborate on the mechanism underlying the priming.

[4] A detailed discussion of the construction of reality is beyond the scope of this paper.

[5] Diagnosticity of inputs is theoretically a highly relevant concept. In the practice of online ratings, however, most of the attribute dimensions in MD rating systems are highly diagnostic for the overall rating because rating platforms have an interest in providing information on important rather than unimportant aspects of the rating object. Nevertheless, diagnosticity could be addressed in future studies.

[6] We use the average of dimensional ratings as a proxy to judge whether the information made accessible through these ratings can be categorized as either positive, negative, or neutral on average. By doing so, we do not intend to hypothesize that a person's weighting of the individual dimensions in the overall rating are necessarily equal.

[7] For the sake of simplicity, the scenario presented uses two versus four dimensions. In practice, the number of dimensions needed to provide a more complete coverage is likely to be higher.

[8] Our studies did not focus on the diagnosticity of the dimensions; thus—by design—we primarily included dimensions high in diagnosticity.

[9] Besides sharing data, preregistration is another important practice to avoid publication bias (Gonzales and Cunningham 2015). Unfortunately, when starting our research, we did not preregister our study; however, because we present not only one but a series of experiments (and present also nonsignificant effects), we believe that our results are both robust and replicable.

[10] We used t-tests for Study 2c; hence, these results are not included in this overview table.

[11] The participants watched *Xiao Xiao*, a two-minute animated martial arts film that has been used in other studies (see, e.g., Schlosser 2005); see e-companion section EC.2 for details.

[12] As such, these dimensions are likely to be highly diagnostic for evaluating restaurants.

[13] We expected a medium-to-large effect size (d = 0.65), suggesting a required sample size of 63 participants per condition (Cohen 1988). Hence, in all studies, we recruited at least 100 participants per condition.

[14] We performed robustness checks using cutoffs of 10, 20, 30, and 40 seconds; our main results were not affected by differences in reading times. We, thus, only report the results for the most conservative cutoff.

[15] In Study 1c, we presented different dimensions; however, the aggregation procedure was the same.

[16] When it comes to random-effects modeling, traditional regression approaches have several drawbacks, including the following: "(a) deficiencies in statistical power related to the problems posed by repeated observations, (b) the lack of a flexible method of dealing with missing data, (c) disparate methods for treating continuous and categorical responses, as well as (d) unprincipled methods of modeling heteroskedasticity and non-spherical error variance (for either participants or items)" (Baayen et al. 2008, p. 391). Linear mixed-effects models can address these drawbacks, especially when dealing with clustered data as is the case in our analyses (see Baayen et al. (2008) for an extensive discussion).

[17] Mixed-effects models rely on assumptions such as linearity, normality of residuals, and homogeneity of variance. Across all experiments, most assumptions were met; however, in some cases, Levene's test showed that homogeneity of variance was violated. In such cases, we reanalyzed our data using unequal-variance models using Bayesian inference, implemented in *R* using *brms* (Bürkner 2017). The results remain qualitatively the same and are available upon request.

[18] We conducted a robustness check to examine whether dimensional ratings can influence overall ratings even if dimensional and overall ratings are elicited on different pages. Using a final sample size of 454 participants recruited using Prolific, we found that overall ratings in the MD condition can differ from those in the SD condition even if dimensional and overall ratings are elicited on subsequent pages. However, this effect was smaller than when both ratings were elicited on the same page (β_1 = .20 versus β_1 = .52) and only marginally significant (p = .054); detailed results are available upon request. This difference is not surprising because memory accessibility (influenced by factors such as time since activation, interfering material, or elaboration and rehearsal) is likely diminished if dimensional and overall ratings are elicited on subsequent pages.

[19] We thank reviewer 3 for this suggestion.

[20] We thank reviewer 3 for this suggestion.

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

[21] These results provide support for the assimilation (carryover) effects discussed by Schwarz and colleagues (e.g., Schwarz and Bless 2007). However, research on part–whole assimilation effects suggests that such effects of general questions on specific questions tend to be smaller than the effects of specific questions on general ones (see, e.g., Schwarz et al. 1991).

[22] In addition, customers in real-world situations typically self-select whether to rate a product/service and on which platform to do so. Although this may influence ratings on each particular platform, we believe that—because of the random assignment of participants to conditions—this should not affect the results of our experimental studies.

[23] Both short films (see e-companion section EC.2) have been used in prior studies (e.g., Schlosser 2005).

[24] The fact that Yahoo! uses these dimensions implies that they are likely to be highly diagnostic for evaluating movies.

[25] In an online pretest, we confirmed that participants indeed rated the short films as expected (details are presented in e-companion section EC.3).

[26] Because we only compared the overall rating scores from two MD conditions (i.e., we did not have clustered data) in Study 2c, we did not follow the analysis strategy used in the other studies (i.e., mixed-effects modeling).

[27] We thank reviewer 3 for this suggestion.

[28] Before rating the university, participants were instructed to watch the film *Xiao Xiao* (used in Study 2b) and to count the stick figures "killed" during the film.

[29] The use of these dimensions by study-advisor.org implies that they are likely to be highly diagnostic for evaluating universities.

[30] Tsekouras (2015, 2017) discussed rating-scale biases such as anchoring the scale at the middle position, presenting a varying number of stars, or adding emotional labels to the scales. Liu et al. (2014) and Ge and Li (2015) also discussed a "dimension rating bias."

[31] We thank reviewer 2 for this suggestion.

[32] We thank reviewer 3 for these suggestions.

[33] We thank reviewer 3 for this suggestion.

References

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Sci.* 57(8):1485–1509.

Ba S, Pavlou PA (2002) Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *Management Inform. Systems Quart.* 26(3):243–268.

Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. *J. Memory Language* 59(4):390–412.

Bates D, Mächler M, Bolker BM, Walker SC (2015) Fitting linear mixed-effects models using lme4. *J. Statist. Software* 67(1):1–51.

Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD (2001) Bad is stronger than good. *Rev. General Psych.* 5(4):323–370.

Bettman JR, Sujan M (1987) Effects of framing on evaluation of comparable and noncomparable alternatives by expert and novice consumers. *J. Consumer Res.* 14(2):141–154.

Bickart B, Schindler RM (2001) Internet forums as influential sources of consumer information. *J. Interactive Marketing* 15(3):31–40.

Bolton GE, Katok E, Ockenfels A (2004) How effective are electronic reputation mechanisms? An experimental investigation. *Management Sci.* 50(11):1587–1602.

Bürkner PC (2017) brms: An R package for Bayesian multilevel models using Stan. *J. Statist. Software* 80(1):1–28.

Burtch G, Hong Y, Bapna R, Griskevicius V (2017) Stimulating online reviews by combining financial incentives and social norms. *Management Sci.* 64(5):2065–2082.

Chen PY, Hong Y, Liu Y (2018) The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. *Management Sci.* 64(10):4629–4647.

Chen Y, Xie J (2005) Third-party product review and firm marketing strategy. *Marketing Sci.* 24(2):218–240.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(3):345–354.

Chintagunta PK, Gopinath S, Venkataraman S (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Sci.* 29(5):944–957.

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Darby MR, Karni E (1973) Free competition and the optimal amount of fraud. *J. Law Econom.* 16(1):67–88.

de Langhe B, Fernbach PM, Lichtenstein DR (2016) Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *J. Consumer Res.* 42(6):817–833.

Decker R, Trusov M (2010) Estimating aggregate consumer preferences from online product reviews. *Internat. J. Res. Marketing* 27(4):293–307.

Dellarocas C (2003) The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Sci.* 49(10):1407–1424.

Dellarocas C (2006) Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Sci.* 52(10):1577–1593.

Dellarocas C, Wood CA (2008) The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Sci.* 54(3):460–476.

Dellarocas C, Zhang XM, Awad NF (2007) Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *J. Interactive Marketing* 21(4):23–45.

Duan W, Gu B, Whinston AB (2008) Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems* 45(4):1007–1016.

Faul F, Erdfelder E, Buchner A, Lang AG (2009) Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* 41(4):1149–1160.

Feldman JM (1992) Constructive processes as a source of context effects in survey research: Explorations in self-generated validity. Schwarz N, Sudman S, eds. *Context Effects in Social and Psychological Research* (Springer, New York), 49–61.

Feldman JM, Lynch JG (1988) Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *J. Appl. Psych.* 73(3):421–435.

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* 19(3): 291–313.

Gao G, Greenwood BN, Agarwal R, McCullough JS (2015) Vocal minority and silent majority: How do online ratings reflect population perceptions of quality? *Management Inform. Systems Quart.* 39(3):565–589.

Ge Y, Li J (2015) Measure and mitigate the dimensional bias in online reviews and ratings. *Proc. Internat. Conf. Inform. Systems* (Fort Worth, TX), 1–11.

Gelman A, Hill J (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, New York).

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Sci.* 31(3):493–520.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Sci.* 23(4):545–560.

Godes D, Silva JC (2012) Sequential and temporal dynamics of online opinion. *Marketing Sci.* 31(3):448–473.

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Goes PB, Lin M, Yeung CA (2014) "Popularity effect" in user-generated content: Evidence from online product reviews. *Inform. Systems Res.* 25(2):222–238.

Gonzales JE, Cunningham CA (2015) The promise of pre-registration in psychological research. https://www.apa.org/science/about/psa/2015/08/pre-registration.

He SX, Bond SD (2013) Word-of-mouth and the forecasting of consumption enjoyment. *J. Consumer Psych.* 23(4):464–482.

Hennig-Thurau T, Gwinner KP, Walsh G, Gremler DD (2004) Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? *J. Interactive Marketing* 18(1):38–52.

Higgins ET (1978) "Saying is believing": Effects of message modification on memory and liking for the person described. *J. Experiment. Soc. Psych.* 14(4):363–378.

Higgins ET, Brendl CM (1995) Accessibility and applicability: Some "activation rules" influencing judgment. *J. Experiment. Soc. Psych.* 31(3):218–243.

Hippler HJ, Schwarz N, Sudman S (1987) *Social Information Processing and Survey Methodology* (Springer, New York).

Ho YC, Wu J, Tan Y (2017) Disconfirmation effect on online rating behavior: A structural model. *Inform. Systems Res.* 28(3):626–642.

Hong YK, Pavlou PA (2014) Product fit uncertainty in online markets: Nature, effects, and antecedents. *Inform. Systems Res.* 25(2):328–344.

Hu N, Pavlou PA, Zhang J (2017) On self-selection biases in online product ratings. *Management Inform. Systems Quart.* 41(2): 449–471.

Hu N, Zhang J, Pavlou PA (2009) Overcoming the J-shaped distribution of product reviews. *Comm. ACM.* 52(10):144–147.

Hu N, Bose I, Gao Y, Liu L (2011) Manipulation in digital word-of-mouth: A reality check for book reviews. *Decision Support Systems* 50(3):627–635.

Hyman HH, Sheatsley PB (1950) The current status of American public opinion. Payne JC, ed. *The Teaching of Contemporary Affairs: Twenty-First Yearbook of the National Council for the Social Studies* (National Education Association, New York), 11–34.

Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. *J. Comput. Graphical Statist.* 5(3):299–314.

Jabine TB, Straff ML, Tanur JM, Tourangeau R (1984) *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines* (The National Academies Press, Washington, DC).

Jiang Y, Guo H (2015) Design of consumer review systems and product pricing. *Inform. Systems Res.* 26(4):714–730.

Kanouse DE, Hanson LR Jr (1987) Negativity in evaluations. Jones EE, Kanouse DE, Kelley HH, Nisbett RE, Valins S, Weiner B, eds. *Attribution: Perceiving the Causes of Behavior* (Lawrence Erlbaum Associates, Hillsdale, NJ), 47–62.

Keller KL (1987) Memory factors in advertising: The effect of advertising retrieval cues on brand evaluations. *J. Consumer Res.* 14(3):316–333.

Kittur A, Chi EH, Suh B (2008) Crowdsourcing user studies with Mechanical Turk. *Proc. 2008 CHI Conf. Human Factors Comput. Systems* (Florence, Italy), 453–456.

Kumar N, Benbasat I (2006) The influence of recommendations and consumer reviews on evaluations of websites. *Inform. Systems Res.* 17(4):425–439.

Kwark Y, Chen J, Raghunathan S (2014) Online product reviews: Implications for retailers and competing manufacturers. *Inform. Systems Res.* 25(1):93–110.

Lee YJ, Hosanagar K, Tan Y (2015) Do I follow my friends or the crowd? Information cascades in online movie ratings. *Management Sci.* 61(9):2241–2258.

Li X, Hitt LM (2008) Self-selection and information role of online product reviews. *Inform. Systems Res.* 19(4):456–474.

Liu Y (2006) Word of mouth for movies: Its dynamics and impact on box office revenue. *J. Marketing* 70(3):74–89.

Liu Y, Chen PY, Hong Y (2014) Value of multi-dimensional rating systems: An information transfer view. *Proc. Internat. Conf. Inform. Systems* (Auckland, New Zealand), 1–18.

Matlin MW, Stang DJ (1978) *The Pollyanna Principle: Selectivity in Language, Memory, and Thought* (Schenkman, Cambridge, UK).

McCloskey H (2015) The 1-2-3's of A/B testing: An intro to split and multivariate tests for product managers. https://community.uservoice.com/blog/ab-split-testing-product.

Moe WW, Schweidel DA (2012) Online product opinions: Incidence, evaluation, and evolution. *Marketing Sci.* 31(3):372–386.

Moe WW, Trusov M (2011) The value of social dynamics in online product ratings forums. *J. Marketing Res.* 48(3):444–456.

Muchnik L, Aral S, Taylor SJ (2013) Social influence bias: A randomized experiment. *Sci.* 341(6146):647–651.

Mudambi SM, Schuff D (2010) What makes a helpful online review? A study of customer reviews on Amazon.com. *Management Inform. Systems Quart.* 34(1):185–200.

Parasuraman A, Zeithaml VA, Berry LL (1988) SERVQUAL—A multiple-item scale for measuring consumer perceptions of service quality. *J. Retailing* 64(1):12–40.

Pavlou PA, Gefen D (2004) Building effective online marketplaces with institution-based trust. *Inform. Systems Res.* 15(1):37–59.

Peterson RA, Wilson WR (1992) Measuring customer satisfaction: Fact and artifact. *J. Acad. Marketing Sci.* 20(1):61–71.

Resnick P, Zeckhauser R, Friedman E, Kuwabara K (2000) Reputation systems. *Comm. ACM.* 43(12):45–48.

Schlosser AE (2005) Posting vs. lurking: Communicating in a multiple audience context. *J. Consumer Res.* 32(2):260–265.

Schuman H, Ludwig J (1982) The norm of even-handedness in surveys as in life. *Amer. Sociol. Rev.* 48(1):112–120.

Schwarz N (1998) Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. *Personality Soc. Psych. Rev.* 2(2):87–99.

Schwarz N (1999) Self-reports: How the questions shape the answers. *Amer. Psych.* 54(2):93–105.

Schwarz N, Bless H (2007) Mental construal processes: The inclusion/exclusion model. Stapel D, Suls J, eds. *Assimilation and Contrast in Social Psychology* (Psychology Press, Philadelphia), 119–141.

Schwarz N, Hippler HJ (1995) Subsequent questions may influence answers to preceding questions in mail surveys. *Public Opinion Quart.* 59(1):93–97.

Schwarz N, Strack F (1991) Context effects in attitude surveys: Applying cognitive theory to social research. *Eur. Rev. Soc. Psych.* 2(1):31–50.

Schwarz N, Strack F, Mai HP (1991) Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quart.* 55(1):3–23.

Sherman SJ, Ahlm K, Berman L, Lynn SJ (1978) Contrast effects and their relationship to subsequent behavior. *J. Experiment. Soc. Psych.* 14(4):340–350.

Simonson I (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. *J. Consumer Res.* 42(6):840–845.

Siroker D, Koomen P (2015) *A/B Testing: The Most Powerful Way to Turn Clicks into Costumers* (John Wiley & Sons, Hoboken, NJ).

Smith A, Anderson M (2016) *Online Shopping and E-commerce.* Technical report, Pew Research Center, Washington, DC.

Srull TK, Wyer RS (1979) The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *J. Personality Soc. Psych.* 37(10): 1660–1672.

Stevens P, Knutson B, Patton M (1995) DINESERV: A tool for measuring service quality in restaurants. *Cornell Hotel Restaurant Admin. Quart.* 36(2):56–60.

Downloaded from informs.org by [62.22.104.105] on 03 July 2023, at 04:12 . For personal use only, all rights reserved.

Sudman S, Bradburn NM, Schwarz N (1996) *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology* (Jossey-Bass, San Francisco).

Sundaram DS, Mitra K, Webster C (1998) Word-of-mouth communications: A motivational analysis. *Adv. Consumer Res.* 25: 527–531.

Surowiecki J (2004) *The Wisdom of Crowds* (Little, Brown and Company, New York).

Thaler RH, Sunstein CR, Balz JP (2012) Choice architecture. Shafir E, ed. *Behavioral Foundations of Public Policy* (University Press Group, West Sussex, UK), 428–439.

Tourangeau R, Rasinski KA (1988) Cognitive processes underlying context effects in attitude measurement. *Psych. Bull.* 103(3): 299–314.

Tsekouras D (2015) Variations on a rating scale: The effect on extreme response tendency in product ratings. *Proc. Eur. Conf. Inform. Systems* (Münster, Germany), 1–16.

Tsekouras D (2017) The effect of rating scale design on extreme response tendency in consumer product ratings. *Internat. J. Electronic Commerce* 21(2):270–296.

Tunc MM, Cavusoglu H, Raghunathan S (2017) Single-dimensional vs. multi-dimensional product ratings in online marketplaces for experience goods. *Proc. Internat. Conf. Inform. Systems* (Seoul, South Korea), 1–16.

Wang A, Zhang M, Hann IH (2018) Socially nudged: A quasi-experimental study of friends' social influence in online product ratings. *Inform. Systems Res.* 29(3):641–655.

Whelan S, Wohlfeil M (2006) Communicating brands through engagement with "lived" experiences. *J. Brand Management* 13(4–5):313–329.

Wyer RS, Srull TK (1986) Human cognition in its social context. *Psych. Rev.* 93(3):322–359.

Zhu F, Zhang XM (2010) Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *J. Marketing* 74(2):133–148.