Behavior Research Methods
https://doi.org/10.3758/s13428-018-1175-8

Decision contamination in the wild: Sequential dependencies in online review ratings

David W. Vinson[1] · Rick Dale[2] · Michael N. Jones[3]

Abstract
Current judgments are systematically biased by prior judgments. Such biases occur in ways that seem to reflect the cognitive system's ability to adapt to statistical regularities within the environment. These cognitive *sequential dependencies* have primarily been evaluated in carefully controlled laboratory experiments. In this study, we used these well-known laboratory findings to guide our analysis of two datasets, consisting of over 2.2 million business review ratings from Yelp and 4.2 million movie and television review ratings from Amazon. We explored how within-reviewer ratings are influenced by previous ratings. Our findings suggest a contrast effect: Current ratings are systematically biased away from prior ratings, and the magnitude of this bias decays over several reviews. This work is couched within a broader program that aims to use well-established laboratory findings to guide our understanding of patterns in naturally occurring and large-scale behavioral data.

Keywords Sequential dependence · Decision making · Data mining · Online reviews · Big data · Cognitive principles

Humans are surprisingly bad at rating the absolute magnitude of their internal cognitive states. Regardless of the task, judgments of the *absolute* magnitude of a stimulus, experience, or feeling are inherently contaminated by *relative* information from the sequence of judgments prior to the current one. Although we tend to believe that our judgment reflects the absolute value of the current experience, a good deal of the judgment is in fact determined by the relative difference between the current experience and the experiences from previous trials (Laming, 1984; Stewart, Brown, & Chater, 2005). This pattern is complicated by the fact that decisions are independently influenced by factors such as stimulus, response, and feedback (see Donkin, Rae, Heathcote, & Brown, 2015, for a review).

These cognitive *sequential dependencies* (SDs) occur whenever behavior on a trial is influenced by behavior on preceding trials. Far from rare, SDs are ubiquitous in cognition, contaminating absolute judgments from low-level perception all the way up to high-level moral judgments. We see the effect of previous trials on the latency, accuracy, and types of errors produced, as well as on the interpretation of ambiguous stimuli. SDs seem to affect all levels of the cognitive system, including motor control (Dixon, McAnsh, & Read, 2012), spatial memory (Freyd & Finke, 1984), face perception (Hsu & Yang, 2013; Liberman, Fischer, & Whitney, 2014), selective attention (Kristjánsson, 2006), decision making (Jesteadt, Luce, & Green, 1977), and language processing (Bock & Griffin, 2000).

SDs have primarily been studied in the laboratory, or at least with well-controlled experimental stimuli. They are more difficult to study in real-world scenarios. Real-world observations are inherently noisy, often occurring sporadically over both time and stimuli, and almost never in isolation. As such, a very large number of trials is often required in order to identify their effects. In this article, we explore SDs in a real-world situation by mining two large, natural datasets of online review ratings from (1) Yelp Inc. and (2) Amazon Inc. We use these datasets to determine whether current review ratings are contaminated by previous reported experiences. First, we will review the SD trends observed in standard laboratory tasks.

SDs in the laboratory

Assimilation occurs whenever the judgment of the stimulus on trial *n* moves closer on the measurement scale to the judgment

✉ David W. Vinson
dave@davevinson.com

[1] University of California, Merced, CA, USA

[2] University of California, Los Angeles, CA, USA

[3] Indiana University, Bloomington, IN, USA

Published online: 02 January 2019

of the stimulus k steps behind, at $n-k$, than it otherwise would have been. *Contrast* is the opposite effect, when the judgment of stimulus n moves farther away on the measurement scale from the judgment of stimulus $n-k$. In this sense, assimilation can be thought of as an attracting force from the preceding stimulus, whereas contrast can be thought of as a repelling force (Zotov, Jones, & Mewhort, 2011).

Much of the early work on SDs was psychophysical in nature and involved rating unidimensional stimuli such as the loudness of a tone or the length of a line (Garner, 1953; Holland & Lockhead, 1968). *Identifying* the absolute magnitude of these stimuli (e.g., line length) has been well studied: Errors when identifying stimulus n assimilate toward the stimulus on trial $n-1$.[1] Participants are more likely, when identifying a stimulus, to estimate its magnitude as being more similar to the preceding stimulus than to identify it as less similar to the preceding stimulus. Oddly, *categorization* of the same stimuli shows the opposite effect—a contrast effect from the previous response. When stimuli are clustered into categories and the response is a category label (e.g., small, medium, large), stimulus n is more likely to be labeled as belonging to a category further away from stimulus $n-1$ on the measurement scale (Stewart, Brown, & Chater, 2002; Ward & Lockhead, 1971).

The contrast effect (repelling) of trial $n-1$ on the category rating of trial n is not limited to low-level perception, but is seen across levels of cognition. As a striking high-level demonstration, consider Parducci's (1968) example of classifying the event of "poisoning a neighbor's barking dog that was bothering you" on a moral judgment scale from 1 to 10 (where 10 is *extremely evil*). This statement was rated as being more evil by participants if it was preceded by a mild judgment ("stealing a towel from a hotel") than if it was preceded by a nastier judgment ("using guns on striking workers")—a contrast effect when classifying moral judgments that mirrors the findings with low-level perceptual stimuli.

This same pattern can be seen in a more recent study by Olivola and Sagara (2009), who found that participants will elect to risk more human lives, as compared to a less risky alternative (with an equal probability of saving the same number of lives), when the number of lives at risk is equal to the probability of the number of lives lost when randomly selecting an observed disaster. The choice is clearly in contrast with one's experiences. Participants are willing to risk more human lives than average when there are a larger number of smaller-casualty events, and less likely to risk more human lives when a higher number of high-casualty events occur. Furthermore, the binary choice (risky vs. sure) decision highlights, importantly, that this cannot be attributed to a scale-interpretation effect (i.e., artifact), a potential criticism of Parducci's (1968) effect. Their findings further emphasize how statistical properties of our environment are reflected in our cognitive system.

Similar patterns of SDs have been seen in a variety of laboratory tasks designed to tap real-world scenarios, including brake initiation latencies in driving behavior (Doshi, Tran, Wilder, Mozer, & Trivedi, 2012), jury evidence interpretation (Furnham, 1986), and clinical assessments (Mumma & Wilson, 1995). In addition, SDs seem to be immune to practice—they are seen even in overlearned and expert behaviors (Doshi et al., 2012).

At first glance, SDs appear to be an irrational bias in decision making (or perhaps in event memory), and traditionally they have been viewed as the natural by-product of low-level brain dynamics, such as residual neural activation. However, more recent theoretical perspectives suggest that SDs may be a rational property of any cognitive system. These accounts characterize SDs in terms of an individual's adaptation to the statistical regularities of a nonstationary environment with related stimulus bundles (Qian & Aslin, 2014; Wilder, Jones, & Mozer, 2010; Yu & Cohen, 2009).

Computational models that explain how SDs emerge from the decision-making process are now being developed, at least for low-level perceptual tasks (e.g., Mozer et al., 2010). These models have great promise, in that they may be reversed and then applied to rating data in order to "decontaminate" the rating, essentially producing a more accurate estimation of the individual's absolute experience of a product or business by removing the pollution from the relative information. As a first step toward decontaminating ratings, our interest is in mining large review datasets such as those from Yelp and Amazon, guided by knowledge from laboratory studies, in order to look for these naturally occurring contaminations that may affect how products and businesses are currently rated by reviewers and can expect to be rated in the future. In the case of Yelp, future business demand is largely influenced by online reviews (Cantallops & Salvi, 2014; Mudambi & Schuff, 2010), which affect a business's revenue between 5% and 9%, with this number increasing by 50% for businesses with more than 50 reviews (Luca, 2011). This has an obvious benefit to the service quality that Yelp and Amazon aim to provide, as well as to providing a more accurate assessment of the products and businesses in question.

In both Yelp and Amazon, reviewers rate their experience with a product or business on a scale of one to five stars. Because both the rating and rating scale are most similar to the features of categorization tasks studied in the laboratory (i.e., what is the best label to classify the exemplar—in this case, experience with the business—on a scale of one to five stars), our predictions were loosely drawn from SDs in categorization. In particular, we expected that within reviewers we would see a *contrast* effect from ratings across products and businesses: For example, an individual's rating would be

[1] Interestingly, the same absolute judgment that assimilates to the most proximal past judgment contrasts with the judgments on stimuli $n-2 \ldots 5$.

artificially inflated if his or her previous rating had been lower, and artificially *deflated* if it had been higher. In this sense, our predictions for review ratings were a simple extension of results from both the perceptual work of Zotov et al. (2011) and the moral judgments of Parducci (1968) or Olivola and Sagara (2009).

Natural datasets are fraught with noise. Yet, what they lack in structure, they make up for in sheer size. We did not anticipate that SDs would play a substantial role in altering the usefulness of user or business ratings on their face. Instead we expected to find *echoes* of these cognitive principles in large datasets of naturally occurring behavior. We consider this work a guided exploration, in an effort to bridge laboratory findings with relevant and functional natural behavior (see Jones, 2017). In a sense, this is somewhat analogous to studying the behavioral patterns of birds in aviary experiments in order to extrapolate to foraging patterns in the wild.

Method

We used two datasets of online reviews, a Yelp, Inc. business review dataset and an Amazon product review dataset (Movies & TV Series[2]). Although business and product reviews are inherently different, they are similar in that users rate their experience with the product or business. Furthermore, Movies & TV Series reviews were selected due to their similarity to Yelp business ratings, in that both ratings occur after intangible experiences, in contrast to tangible products. Although we do not think tangible product reviews would lack SD effects, this was not tested. We used the most recent version of the Yelp Inc. dataset at the time of the research ("Round Seven"), which was released as part of Yelp's dataset challenge.[3] The dataset we used consists of just over 2.2 million reviews spanning 12 years, from 2004–2016, with ratings from one (negative) to five (positive) stars (μ = 3.76), with approximately 552,000 reviewers rating roughly 77,000 businesses. Reviews were provided from nine cities (Edinburgh, Montreal, Karlsruhe, Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, and Madison) across four countries (the United States, Canada, Scotland, and Germany). The Yelp review data are organized in a data format referred to as JSON ("JavaScript Object Notation"), in which each line consists of a single JSON entry, for a user, review, and so forth. For our present analysis, we extracted the Yelp user's unique identifier, their star rating, and the time stamp of that rating. We then ordered the data by reviewer and date for further analysis. Star ratings follow a J-shaped distribution, with mostly four- and five-star ratings, a dip in two-star ratings, and roughly equal numbers of one- and three-star ratings (see Hu, Zhang, & Pavlou, 2009, for a further review of such distributions). The number of reviews increased steadily over Yelp's lifetime, consistent with Moore's law. Similarly, the Amazon product review dataset—consisting of just over 4.6 million reviews dating as far back as 1998, only four years after the site's inception in 1994, and up to 2015—also shows a J-shaped distribution over star ratings (μ = 4.19) and increasingly more reviews over time. The downloaded Amazon dataset is organized in a CSV file consisting of the user's unique identifier, the item's unique identifier, a star rating, and the time stamp.

In both the Yelp and Amazon datasets, reviews occasionally occur at the same time (i.e., have the same time stamp). That is, even after organizing the data, there will be some inherent "noise" in our analysis, since reviews that consist of the same time stamp by a single reviewer are randomly organized. This sort of noise is natural within larger datasets. However, due to the size of each dataset, if there was a true SD effect, our analyses should not be affected by this noise. The distributions for both Yelp and Amazon can be seen in Fig. 1.

We tested whether previous review ratings influenced the current rating within a user. If an individual's current review were sequentially dependent on the previous review, it would likely be repelled from previous reviews, showing a contrast effect (cf. Zotov et al., 2011). We anticipated that these effects would dissipate, the farther away the previous review was from the current review. One possibility, and something we will later address in the Discussion section, is the development of a measure of "bias" or "deviation" for each product/service by comparing a given review received for a given product to the average review received for the same product that was preceded by a median review (e.g., four-star reviews).

Measures

We first calculated the deviation of the current review rating from its mean:

$$R_x – M(R_{T–x}),$$

where R_x is the current rating and $M(R_{T–x})$ is the average rating for the reviewer with the current value x removed. The score is thus the deviation of the current rating from the average of the prior ratings. This means that for a given user, the mean is dynamically adjusted over time. We used a mean adjustment, without standardization, because it can be interpreted in the original scale of the star ratings. In this way, for example, we can interpret + 0.3 as 30% of the way to an entire increment of one "star" value. After obtaining this deviation score, we

[2] We extracted this dataset from https://snap.stanford.edu/data/web-Amazon.html "movies & TV reviews." Further information can be found on this website, including a recent update, but obtaining the data requires emailing Julian McAuley (julian.mcauley@gmail.com) to obtain a link.

[3] Further information on how to access the dataset for free as part of Yelp's dataset challenge can be found at www.yelp.com/dataset_challenge.

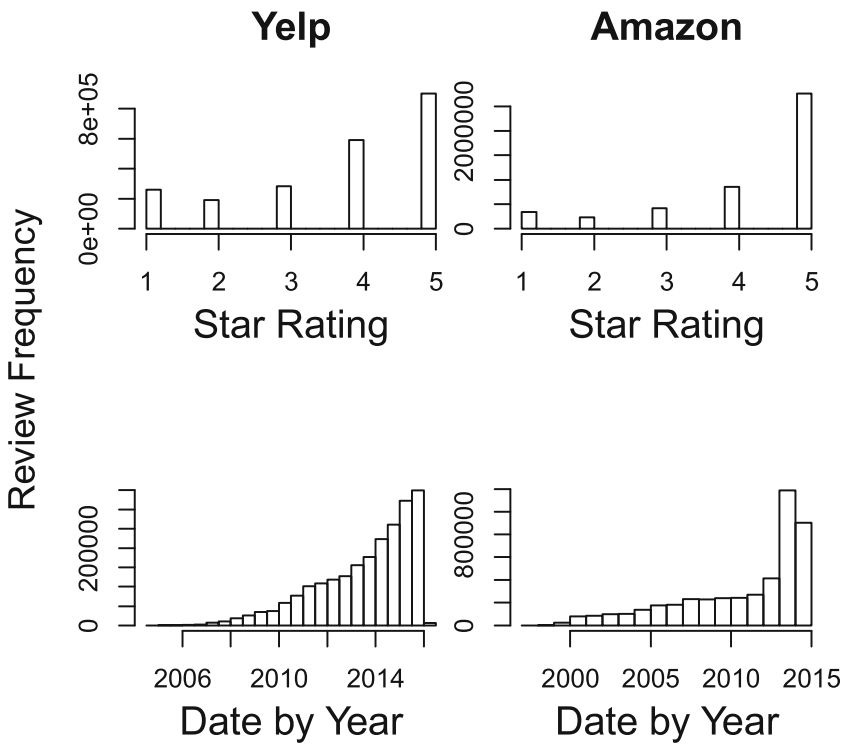

Fig. 1 (Top left) Frequency of reviews by star rating (Yelp). (Bottom left) Frequency of reviews by year (Yelp). (Top right) Frequency of reviews by star rating (Amazon). (Bottom right) Frequency of reviews by year (Amazon)

centered the values. It is important to note that the results are unlikely to have been influenced by this centering, since the standardization into z scores should have only translated the values linearly, leaving the statistical patterns unchanged. This allowed us to determine directly whether the reviewer's current rating was systematically biased away from his or her average response relative to the value of the preceding $n - k$ review(s). To assess how distance is related to this deviation measure, we used *review distance* (k), an ordinal lag measure of the number of reviews (k) between the current review and the previous review.

Results

Yelp

We first determined whether a reviewer's current review was related to his or her previous review rating at distance k for Yelp reviews.[4] Figure 2 presents the mean, with standard error bars, for the deviation of the current review rating from the mean (y-axis) by the previous star ratings (x-axis) at seven different review distances (k) within Yelp reviewers. The figure reveals an asymmetric *contrast* effect that dissipates the farther away the previous review is from the current review. At $n - 1$ (the immediately preceding review), for example, a one-star rating resulted in an increase in the subsequent rating from the overall mean rating. The opposite was the case if the $n - 1$ rating was five stars—the subsequent rating deviated toward a lower star rating relative to the average prior review ratings. In this sense, the data are very much consistent with Olivola and Sagara's (2009) experiment, in that the current rating is systematically biased in the opposite direction from previous experience.

To assess these results quantitatively, we used eight linear models to predict the current review ratings by the $n - k$ ratings for each of seven different values of k and by a random review baseline. To create this baseline, we treated each value of k as distinct, thereby shuffling the results within reviewers. A crucial reason for the use of a random review baseline was to show that our observed effects are not driven by regression to the mean. Although the effect of the previous review on the current review deviation does fall away at each distance of k, our interest was in the deviation of the current review from the mean, not in the mean itself. For this baseline, the expected value of the deviation of the current review from the mean was zero at any distance of k. If our results were driven by a regression to the mean, we would have seen a similar pattern at each distance k to that of the baseline. Aside from reasons of computational and interpretive simplicity of the linear regression, there were two additional reasons we employed this

[4] All code used to visualize and analyze our results can be found here: https://github.com/DaveVinson/sequential-dependence-reviews. Yelp's data agreement does not permit sharing the data publicly; however, if contacted directly, the authors can share the exact dataset uses for these analyses (because Yelp regularly adds to its challenge dataset).

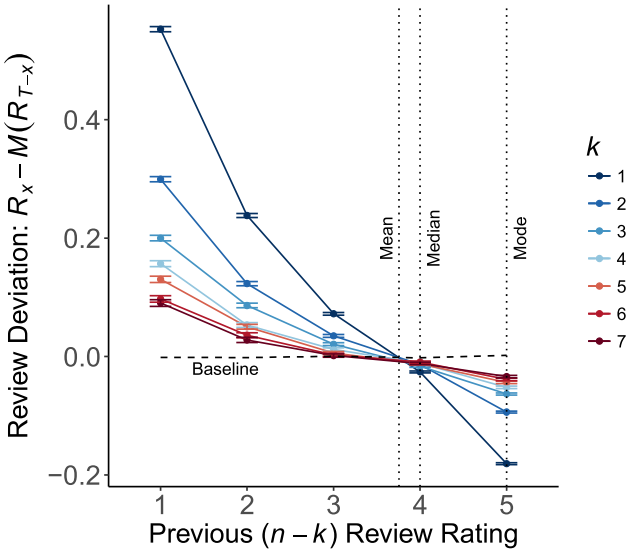

Fig. 2 Deviation of the current review rating from the reviewer's average rating (*y*-axis) in relation to the previous review rating (*x*-axis) at a distance of *k* reviews, for Yelp reviews

statistical method. First, the size of our dataset was quite large, including many thousands of individual reviewers; it was thus unlikely that the nonindependence of some observations would impact the results. Second, we had straightforward linear hypotheses about the observed contrast effect, seen in Fig. 2. In this way, we were able to assess the relative linear impact that each value of *k* had on the current review rating. We calculated this value in R using lm($[R_x - M(R_{T-x})] \sim k$). The results, presented in Table 1, reveal that as the value of *k* increases, as the current review is displaced farther from the previous review, the contrast effect dissipates. A randomly resampled review baseline, in which all reviews were first shuffled and then used to predict the reviewer's current rating ("Baseline" in Table 1), showed no significant effect on the current review rating. With the exception of the random review baseline, all values of *k* showed a significant negative relationship with current review ratings, accounting for ~ 2% of the variance at the closest review distance (k = 1). The

Table 1 Regression model for *k* distances by Yelp reviewer

k	99.9% *CI*	*F* (*df*)	R^2_{adj}
Baseline	(– .001, .003)	0.40 (1, 1.9×10^6)	< .00001
1	(– .17, – .16)	4.5×10^4 (1, 1.7×10^6)	.03
2	(– .09, – .08)	1.1×10^4 (1, 1.4×10^6)	.008
3	(– .06, – .06)	4,376 (1, 1.4×10^6)	.004
4	(– .05, – .04)	2,282 (1, 1.1×10^6)	.002
5	(– .04, – .03)	1,477 (1, 1.0×10^6)	.001
6	(– .03, – .03)	829 (1, 9.4×10^5)	.001
7	(– .03, – .02)	610 (1, 8.7×10^5)	< .001

CI is the 99.9% confidence interval, and *df*, is the residual degrees of freedom, equal to the number of observations for each *k* value

residual degrees of freedom, in the *F* column of the table, are equal to the number of observations in that category. At k = 1, the number of observations was 2.2 × 10^6 = 2.2 million reviews.

To determine whether there was systematic decay in the effect of previous review ratings at different distances of *k*, we first subtracted the review value at each *k* distance from the current estimated review rating, and then squared this value:

$$[R_x - M(R_{T-x})]^2.$$

We treated *k* as a continuous variable and used it to predict $[R_x - M(R_{T-x})]^2$. There was a significant negative relationship between *k* and the magnitude of its effect on the current review, $F(1, 8.2\times10^6) = 2.7 \times 10^4$, R^2 = .003, *CI* = (– .067, – .064), $p < .001$, such that as *k* increases, the magnitude of the effect of the previous review decreases (Fig. 3).

Amazon

The contrast effect found for within-reviewer Yelp ratings was replicated in the Amazon review ratings (Fig. 4). As with Yelp, the contrast effect dissipated, the farther away the previous review was from the current review. The results, presented in Table 2, reveal that as the value of *k* increases the contrast effect dissipates. With the exception of the random review baseline, all values of *k* show a significant negative relationship with current review ratings, accounting for ~ 1.5% of variance at the closest review distance (k = 1). At k = 1, the number of observations was 4.6 million reviews.

Again, we treated *k* as a continuous variable and predicted the magnitude of the observed contrast effect. There was a significant negative relationship between *k* and the magnitude of its effect on the current review, $F(1, 1.1 \times 10^7) = 1.3 \times 10^4$,

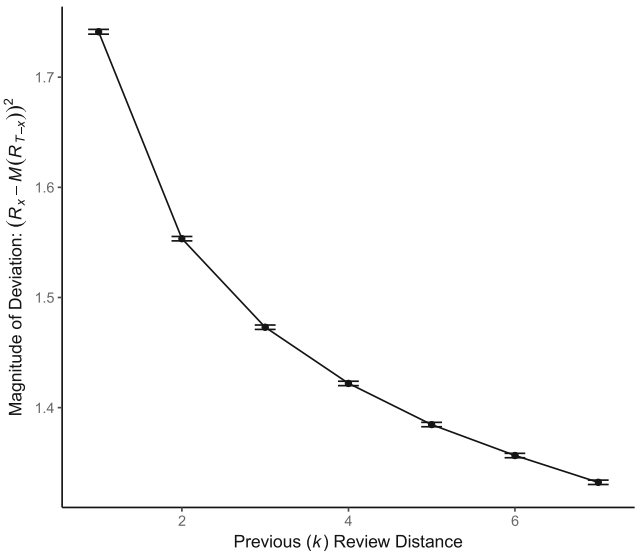

Fig. 3 Magnitude of the contrast effect on the current review at *k* distance for Yelp reviews

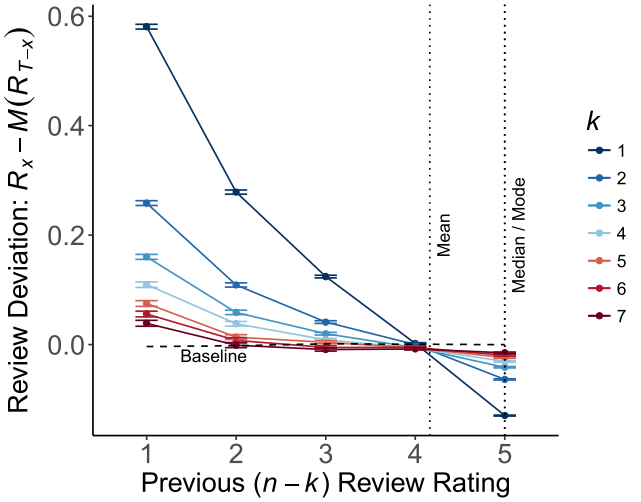

Fig. 4 Deviation of the current review rating from the reviewer's average rating (*y*-axis) in relation to the previous review rating (*x*-axis) at a distance of *k* reviews, for Amazon reviews

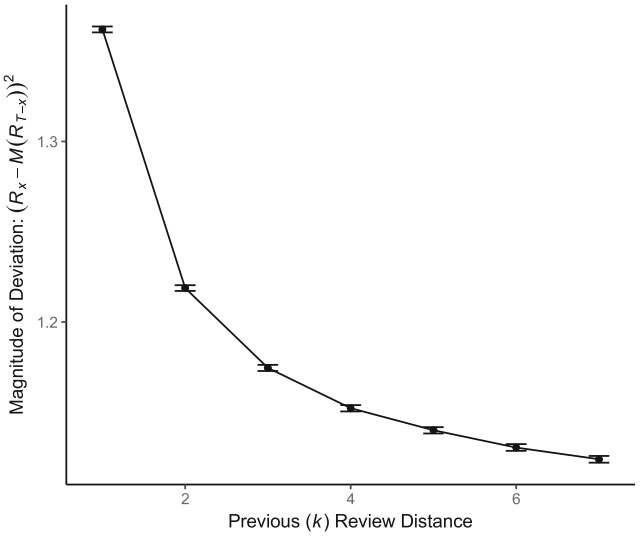

Fig. 5 Magnitude of the deviation of the current review rating from the reviewer's average relative to the previous review rating at distance *k* for Amazon reviews

$R^2 = .001$, $CI = (-.04, -.04)$, $p < .001$, such that as *k* became larger, the magnitude of the effect of the previous review decreased (Fig. 5).

Discussion

We evaluated the presence of SDs in online review ratings across two different platforms and found subtle but significant dependencies. In fact, while the SD effects are indeed small in each set, they are astonishingly similar for both Yelp and Amazon. They support the predictions guided by laboratory experiments in both categorization tasks (Zotov et al., 2011) and moral judgments (Olivola & Sagara, 2009). In both online review datasets, the current ratings deviated from the average rating in contrast to preceding ratings: If a reviewer's previous rating was positive, the current rating was more likely to be less positive than the average rating. In addition, the contrast pattern was asymmetric. For example, a poor experience (one-star rating) with a restaurant made the subsequent restaurant appear more positive, and this upward contrast was more powerful than the downward impression of a restaurant following an excellent preceding experience (five stars). However, this asymmetry may be the result of a positivity bias in reviews. The "even point" in our data appears to be four-star reviews, where a prior review does not bias it. If this were centered on three stars, the observed contrast effect might appear more symmetric. Furthermore, the effect dissipates, the farther the previous review is from the current review. We found no effect of previous review ratings when reviews were randomly ordered. The findings from our study suggest that the observed contrast effects may be stable across reviewer contexts and products.

Table 2 Regression model for *k* distances by Amazon reviewer

k	99.9% *CI*	*F* (*df*)	R^2_{adj}
Baseline	(– .001, .002)	1.36 (1, 3.1×10^6)	< .00001
1	(– .16, – .16)	6.84×10^4 (1, 2.5×10^6)	.026
2	(– .07, – .07)	1.08×10^4 (1, 1.9×10^6)	.006
3	(– .04, – .04)	3,419 (1, 1.6×10^6)	.002
4	(– .03, – .03)	1,423 (1, 1.4×10^6)	.001
5	(– .02, – .02)	573 (1, 1.2×10^6)	< .001
6	(– .02, – .01)	277 (1, 1.1×10^6)	< .001
7	(– .01, – .01)	105 (1, 1.1×10^6)	< .001

CI is 99.9% confidence interval and *df*, the residual degrees of freedom equal to the number of observations for each *k* value

Note that the Yelp and Amazon review distributions are not normal, exhibiting a J shape, or a bimodal distribution at one star and four to five stars, with a mean of 3.75. Recent studies have suggested that a J-shape bimodal distribution, unique to review data, may be the result of an underreporting bias (Hu et al., 2009), such that reviewers are more likely not to provide reviews when the average business rating is similar to their own experience. Interestingly, critics are more likely to have a unimodal distribution, whereas noncritic reviewers tend to produce a J-shaped distribution (Dellarocas & Narayan, 2006).

Computational models that explain how sequential dependencies emerge from the decision-making process can help decontaminate current evaluations in order to produce a more accurate measure of one's experience (e.g., Mozer et al., 2010). Such models, though currently developed only for low-level perceptual tasks, might be fruitfully applied to areas such as online rating systems that are shown to impact a business's future success (Luca, 2011). Our present work is a step toward uncovering contamination effects that may be a rational property of the cognitive system within naturally

occurring behavior. Developing psychologically constrained tools that can adjust for such effects might help provide ratings that better reflect a consumer's experience.

The Yelp and Amazon datasets indeed contain sources of noise. Reviewers sometimes don't review for various reasons, and businesses change their names and their products adapt in real time to the demands of consumer behavior. SD experiments in the laboratory have control over the stimulus presented on each trial, but we lose this experimental control in the real world. Hence, there are several interpretations that we cannot rule out. For example, following a terrible experience with a restaurant, a rater may pay more attention to the selection process, and as a consequence may actually go to a better restaurant the next time. The observed sequential dependency may be a change in selection behavior rather than a bias in decisions. This is particularly why we believe that the approaches of laboratory experimentation and real-world data mining complement each other so well. The cognitive mechanism can be "captured" and studied in the controlled setting of the lab, and then "released" back into the real world, and we have reasonably good indications of what kinds of patterns to then search for in big data that will reveal echoes of the cognitive system operating in the wild.

Research that tests laboratory-derived cognitive principles in the wild could also expand our understanding of these principles. The results here reveal that the effect size is smaller than might be expected, but the results also support the broader interest in SDs. These dependencies unequivocally appear in the noisy and nonstationary environment of human experience, in natural contexts such as evaluative activities during consumption (e.g., of news; Olivola & Sagara, 2009). But the promise of this research goes beyond making the cognitive principles relevant to daily life. The data resources supplied by Amazon and Yelp will allow us to test the boundaries of these cognitive principles. Are there contexts in which SDs are weakened, or even enhanced? For example, one-off experiences on Yelp, such as highly expensive restaurants, may be encoded in human memory quite differently from a restaurant experience that one might expect to have on a regular basis. These one-off experiences are unlikely to obliterate the SD effect, but they may be encoded in memory differently, and we might predict that they would show somewhat weakened SDs. Such questions, which are outside the scope of the present article, may be tested by examining connections among variables in the generous array of information afforded by these natural datasets. Future investigation of the structure of these datasets will allow researchers to search for these boundaries, and thus to refine cognitive theory.

It is worth emphasizing that using cognitive principles with natural data can help practical endeavors in industry. Industry has become extremely focused on the importance of automated recommendation engines based on machine learning. These systems affect every facet of our daily lives, helping us select options on the basis of our previous preferences and global preferences across all individuals. But it is important to note that the upper limit on how well machine-learning systems can perform is dictated by the quality of the data provided by humans. In our example, the human raters do not provide unbiased data from their experiences with the product or company. What is worse, they are probably not aware of this bias. Basing a recommendation system on sequentially contaminated data will be less than optimal, but for reasons that are not random. Hence, we reiterate the potential importance of the past century of experiments in psychological science to modern data-mining enterprises.

Author note This work was funded by NSF BCS-1056744 to M.N.J. D.W.V. was supported by an IBM PhD fellowship.

References

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General*, *129*, 177–192. https://doi.org/10.1037/0096-3445.129.2.177

Cantallops, A. S., & Salvi, F. (2014). New consumer behavior: A review of research on eWOM and hotels. *International Journal of Hospitality Management*, *36*, 41–51. https://doi.org/10.1016/j.ijhm.2013.08.007

Dellarocas, C., & Narayan, R. (2006). A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth. *Statistical Science*, *21*, 277–285.

Dixon, P, McAnsh, S., & Read, L. (2012). Repetition effects in grasping. *Canadian Journal of Experimental Psychology*, *66*, 1–17.

Donkin, C., Rae, B., Heathcote, A., & Brown, S. D. (2015). Why is accurately labelling simple magnitudes so hard? A past, present and future look at simple perceptual judgment. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 121–141). Oxford, UK: Oxford University Press.

Doshi, A., Tran, C., Wilder, M. H., Mozer, M. C., & Trivedi, M. M. (2012). Sequential dependencies in driving. *Cognitive Science*, *36*, 948–963.

Freyd, J. J., & Finke, R. A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 126–132. https://doi.org/10.1037/0278-7393.10.1.126

Furnham, A. (1986). The robustness of the recency effect: Studies using legal evidence. *Journal of General Psychology*, *113*, 351–357.

Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. *Journal of Experimental Psychology*, *46*, 373–380. https://doi.org/10.1037/h0063212

Holland, M. K., & Lockhead, G. R. (1968). Sequential effects in absolute judgments of loudness. *Perception & Psychophysics*, *3*, 409–414. https://doi.org/10.3758/BF03205747

Hsu, S.-M., & Yang, L.-X. (2013). Sequential effects in facial expression categorization. *Emotion*, *13*, 573–586. https://doi.org/10.1037/a0027285

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*, *52*, 144–147. https://doi.org/10.1145/1562764.1562800

Jesteadt, W., Luce, R. D., & Green, D. M. (1977). Sequential effects in judgments of loudness. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 92–104. https://doi.org/10.1037/0096-1523.3.1.92

Jones, M. N. (2017). *Big data in cognitive science*. New York, NY: Taylor & Francis.

Kristjánsson, Á. (2006). Simultaneous priming along multiple feature dimensions in a visual search task. *Vision Research*, *46*, 2554–2570.

Laming, D. (1984). The relativity of "absolute" judgements. *British Journal of Mathematical and Statistical Psychology*, *37*, 152–183.

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Current Biology*, *24*, 2569–2574. https://doi.org/10.1016/j.cub.2014.09.025

Luca, M. (2011, September 16). *Reviews, reputation, and revenue: The case of Yelp.com* (Harvard Business School NOM Unit Working Paper, 12-016). Cambridge, MA: Harvard University, School of Business.

Mozer, M. C., Pashler, H., Wilder, M., Lindsey, R. V., Jones, M. C., & Jones, M. N. (2010). Decontaminating human judgments by removing sequential dependencies. In J. Laffterty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culota (Eds.), *Advances in neural information processing systems 23* (pp. 1705–1713). La Jolla, CA: NIPS Foundation.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon. com. *MIS Quarterly*, *34*, 185–200.

Mumma, G. H., & Wilson, S. B. (1995). Procedural debiasing of primacy/anchoring effects in clinical-like judgments. *Journal of Clinical Psychology*, *51*, 841–853.

Olivola, C. Y., & Sagara, N. (2009). Distributions of observed death tolls govern sensitivity to human fatalities. *Proceedings of the National Academy of Sciences*, *106*, 22151–22156.

Parducci, A. (1968). The relativism of absolute judgments. *Scientific American*, *219*, 84–90. https://doi.org/10.1038/scientificamerican1268-84

Qian, T., & Aslin, R. N. (2014). Learning bundles of stimuli renders stimulus order as a cue, not a confound. *Proceedings of the National Academy of Sciences*, *111*, 14400–14405.

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 3–11. https://doi.org/10.1037/0278-7393.28.1.3

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*, 881–911. https://doi.org/10.1037/0033-295X.112.4.881

Ward, L. M., & Lockhead, G. R. (1971). Response system processes in absolute judgment. *Perception & Psychophysics*, *9*, 73–78. https://doi.org/10.3758/BF03213031

Wilder, M., Jones, M., & Mozer, M. C. (2010). Sequential effects reflect parallel learning of multiple environmental regularities. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 2053–2061). La Jolla, CA: NIPS Foundation.

Yu, A. J., & Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? In *Advances in neural information processing systems 21* (pp. 1873–1880). La Jolla, CA: NIPS Foundation.

Zotov, V., Jones, M. N., & Mewhort, D. J. (2011). Contrast and assimilation in categorization and exemplar production. *Attention, Perception, & Psychophysics*, *73*, 621–639. https://doi.org/10.3758/s13414-010-0036-z