Expressions of doubt and trust in online user reviews

Evans, Anthony; Stavrova, Olga; Rosenbusch, Hannes

Published in:
Computers in Human Behavior

DOI:
10.1016/j.chb.2020.106556

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Computers in Human Behavior 114 (2021) 106556

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: http://www.elsevier.com/locate/comphumbeh

Full length article

Expressions of doubt and trust in online user reviews

Anthony M. Evans *, Olga Stavrova, Hannes Rosenbusch

Department of Social Psychology, Tilburg University, P.O. Box 90153, 5000LE, Tilburg, Netherlands

ARTICLE INFO

Keywords:
Trust
Doubt
Confidence
Online reviews
Advice-taking

ABSTRACT

How do expressions of doubt affect trust in online reviews? Some previous studies find that people trust confident advisors more than doubtful advisors, whereas others find doubtful advisors are trusted more. We tested the effects of expressing doubt using Yelp data and in a controlled experiment: In Study 1, reviews from Yelp (N = 5.9 million) were coded using the Linguistic Inquiry Word Count (LIWC) software. Reviews with doubtful language were seen as more useful, and this result was robust when controlling for other psychological and linguistic variables. In Study 2, participants (N = 660) evaluated reviews with doubtful or confident conclusions; doubtful reviews were seen as more likely to be written by actual consumers. In both studies, the positive effects of doubt were stronger for positive (vs. negative) reviews, suggesting doubt mitigates concerns about fake positive reviews. The present study emphasizes the advantages of expressing doubt.

1. Expressions of doubt and trust in online user reviews

Online user reviews play an important role in consumer decision-making (De Langhe, Fernbach, & Lichtenstein, 2015). The majority of U.S. American adults have experience with reading (82%) and writing (61%) online reviews (Pew Research Center, 2016). However, online reviews often present readers with contradictory advice or opinions. As a result, trust in online reviews is a major issue, and many readers (48%) find it difficult to determine the truthfulness of reviews (Pew Research Center, 2016). When people are presented with contradictory reviews, what cues do they use to decide which reviews are trustworthy?

The present research investigates how expressions of doubt influence trust in online reviews. Previous work on the role of doubt in advice-giving has led to conflicting results, with some studies finding that people trust confident advisors (Thomas & McFadyen, 1995), whereas other studies find that people prefer advice from doubtful advisors, especially in environments where advice-givers may have ulterior motives or conflicts-of-interest (Van de Calseyde, Keren, & Zeelenberg, 2014). We add to this literature by using observational data from Yelp and an experiment to examine how doubtful language influences the perceived trustworthiness of online reviews.

1.1. Trust in online user reviews

A range of factors, such as product types (Sen & Lerman, 2007) and reader characteristics (Filieri, Alguezaui, & McLeay, 2015) shape whether people trust the opinions expressed in online user reviews. Among these factors, perceptions of author trustworthiness are important (Sen & Lerman, 2007), and perceived trustworthiness is especially relevant when consumers make low-involvement purchases (Filieri, Hofacker, & Alguezaui, 2018). More specifically, readers are concerned with the authenticity of overly positive reviews (De Langhe et al., 2015; Mukherjee, Liu, & Glance, 2012). Many positive reviews are fraudulent, either provided by paid (crowdsourced) authors or written by employees and business owners (Luca & Zervas, 2016).

What cues do readers use to judge the trustworthiness of online reviews (and specifically, positive reviews)? Recent studies have found that users focus on cues that signal the thoughtfulness and cognitive effort of reviewers (Filieri et al., 2018; Kim & Gupta, 2012). For example, Kupor & Tormala (2018) found that readers are more likely to trust reviews that deviate in tone from the default opinion; and this effect occurs because deviating reviews are seen to be more thoughtful (Kupor & Tormala, 2018). Hence, if there are many extremely positive reviews, then moderately positive reviews are trusted more (and vice versa).

User trust is also influenced by the language used by reviewers. Studies on the textual cues present in reviews have examined how the expression of different emotions also influence trust: Users are less likely to trust reviews expressing strong negative emotions, as they are attributed the expressed emotions to the irrationality of the reviewer, rather than the poor quality of the reviewed product (Kim & Gupta, 2012; Lee, Jeong, & Lee, 2017). More specifically, angry (rather than

* Corresponding author.
E-mail address: A.M.Evans@uvt.nl (A.M. Evans).

https://doi.org/10.1016/j.chb.2020.106556
Received 16 January 2020; Received in revised form 12 June 2020; Accepted 7 September 2020
Available online 7 September 2020

anxious) language negatively impacts perceptions of the reviewer (Yin, Bond, & Zhang, 2014). Note that the effects of negative emotions disappear when there is consensus across many different reviews, in which case the reviewer is no longer judged poorly for expressing negative emotions (Kim & Gupta, 2012).

Moving beyond the effects of specific emotions, it is important to consider other textual cues that might influence perceptions of reviewer trustworthiness. We introduce the idea that expressions of doubt play a role in how readers perceive the authors of reviews.

1.2. Perceptions of doubt

Doubt is defined as "the subjective uncertainty that people experience when assessing the correctness of their decisions, beliefs, or opinions" (van de Calseyde, Zeelenberg, & Evers, 2018). Expressions of doubt (or the lack thereof) play an important role in advice-taking (Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000). According to the confidence heuristic, people generally prefer advice from confident, rather than doubtful, sources. In one study, Price and Stone (2004) presented participants with information about two financial advisors. One of the advisors made extremely confident predictions (predicting that a stock will increase in value with 90% confidence), while the other advisor made moderately confident predictions (predicting the same outcome with 60% confidence). When both advisors were equally accurate, most participants preferred the extremely confident advisor. This preference was driven by the assumption that the more confident advisor was also more knowledgeable.

Building on these results, Gaertig and Simmons (2018) drew an important distinction between reactions to general confidence (e.g., the extent to which advisors are certain or doubtful of their given advice) versus the uncertainty of the advice itself (e.g., the specificity of an advisor's prediction). Consistent with the confidence heuristic, Gaertig and Simmons (2018) found that people disliked advisors who were doubtful of their own opinions; participants reacted negatively to advisors who admitted that they were unsure of their own advice. However, participants did not necessarily dislike uncertain advice. In other words, forecasters who predicted certain outcomes ("the weather tomorrow will be 20 °C″) were not preferred over forecasters who predicted uncertain outcomes ("the weather tomorrow will be between 10 °C and 30 °C″) (Gaertig & Simmons, 2018). These findings show that it is important to consider how authors express doubt in their advice.

Additional studies found that advisor confidence becomes less important when people have access to other types of information: Notably, advisor accuracy is more important than advisor confidence (Sah, Moore, & MacCoun, 2013; Tenney, MacCoun, Spellman, & Hastie, 2007). When no accuracy information is available, people judge advisors primarily based on their confidence. However, when information about accuracy is provided, advice-takers strongly prefer accurate (vs. inaccurate) advisors (Tenney et al., 2007). Similarly, people are also more willing to accept the recommendations of doubtful advisors when they make optimistic (vs. pessimistic) predictions (Stavrova & Evans, 2019). Advisees no longer prefer confident advisors when these advisors make pessimistic forecasts about the future. Doubt is only one several cues that people use to evaluate the quality of advice.

In the following section, we consider how doubt may influence user trust in online consumer reviews.

2. Expressions of doubt in consumer reviews

We examine how the use of doubtful (vs. confident) language influences trust in online reviews. Although the confidence heuristic suggests that people generally prefer confident (vs. doubtful) advisors (Price & Stone, 2004), there is also good reason to expect that users in online environments may place greater trust in doubtful (vs. confident) user reviews. Importantly, online review platforms are low-trust environments, as users typically cannot judge reviewers based on their prior behavior and have little-to-no information about reviewers' identities (Evans & Krueger, 2016). Under these conditions, users have valid concerns about the presence of fake reviews (Luca & Zervas, 2016; Mukherjee et al., 2012). In turn, this general lack of trust may influence how users perceive expressions of doubt.

Some prior research suggests that doubt is perceived as a signal of honesty in low-trust environments, where advice-givers may have ulterior motives or salient conflicts-of-interest (Van de Calseyde et al., 2014). In one study, participants read about two mechanics who both confirmed that they could solve an automotive problem: one mechanic responded quickly, the other mechanic responded slowly. In this study, participants always perceived the slower mechanic as more doubtful. However, the consequences of expressing doubt depended on the situation. In a high-trust situation, where both mechanics were financially secure, participants preferred the mechanic who responded more quickly because they believed that the fast mechanic was more confident and more competent. However, in a low-trust situation, where both mechanics were near bankruptcy (and hence, desperately needed additional business), participants preferred the slower mechanic. When the mechanics faced a salient conflict-of-interest (i.e., they would go bankrupt without recruiting additional clients), participants chose the slower (and more doubtful) mechanic, believing that the slow mechanic was more honest.

Given users' trust issues with online review platforms, we hypothesize that doubtful reviews are trusted more than confident reviews. Furthermore, we predict that the positive effects of doubt are strongest when users have heightened concerns about the authenticity of reviews. Previous work indicated that reviewers are concerned with the trustworthiness of uncritical, overly positive reviews (Kupor & Tormala, 2018; Luca & Zervas, 2016; Mukherjee et al., 2012), as these reviews may come from dubious sources (e.g., business owners may post reviews themselves, or use crowdsourcing to generate fake positive reviews). Therefore, we hypothesize that expressions of doubt would have a stronger positive effect on trustworthiness for positive (vs. negative) reviews.[1]

3. Overview of studies

We report two studies: In Study 1, we examine the effects of doubtful language in a dataset from Yelp.com. We asked if reviews that contained doubtful (vs. confident) words were more likely to receive useful votes from users. In Study 2, we conducted an experiment where participants were presented with a standardized set of reviews with doubtful or confident language. Participants rated the authenticity of each review (i.e., the likelihood that the review was written by an actual consumer). We hypothesized that doubtful reviews would be trusted more than confident reviews, and that the positive effects of doubt would be stronger for positive (vs. negative) reviews. Data and syntax to reproduce our analyses are available at https://osf.io/w5dy7/?view_only=420062c267084a5fbf1c16146b8da3df.[2] The following studies were approved by the Ethics Review Board at our host institution.

3.1. Study 1

We conducted text analyses using the Yelp Open dataset, a publicly available dataset which contains over 5.9 million reviews. Our primary outcome of interest was the number of "useful" votes that reviews

[1] Readers may also have reason to distrust some negative reviews written with ulterior motives (for example, reviews written by vengeful customers or business competitors). However, questionable negative reviews are much less common than questionable positive reviews (Luca & Zervas, 2016).

[2] Sharing review text would violate the terms of service for use of the dataset. However, the reviews can be downloaded directly from Yelp and matched with our dataset using the 'review_id' variable.

received from readers; on Yelp, readers can vote to indicate that reviews are useful, funny, and/or cool. Each user can potentially provide one vote in each of the three categories. We focused on usefulness as our primary outcome because it was the most closely associated to the concept of review trust. To measure doubtfulness, we processed review texts using the Linguistic Inquiry Word Count software (Pennebaker, Boyd, Jordan, & Blackburn, 2015), creating a measure to capture how often each review contained words related to two linguistic categories reflecting doubt, tentativeness and (lack of) certainty.

3.1.1. Method

Dataset. We analyzed reviews from the 2017 Yelp Open Dataset (Yelp, 2017). This dataset consists of 5,996,996 reviews from 188,593 businesses, written by 1,518,169 unique users. The reviews were selected from businesses in 1111 North American cities and were published from the period of October 12, 2004 to July 2, 2018. Two reviews were excluded from our analyses: one contained an impossible number of usefulness votes (−1) and the other contained and an impossible star rating (0). These values are likely the result of coding errors in the raw data file, as Yelp does not allow users to downvote reviews and star ratings can only be made on a scale from 1 to 5.

Data Processing. The raw data file was provided by Yelp as a JSON database, which was converted into CSV files containing the information of reviews, businesses, and users (Butler, 2018). These files were then merged into one dataset.

Useful votes. Our primary dependent variable was the total number of times readers on Yelp voted published reviews as useful ($M = 1.37$, $SD = 3.71$). The distribution of usefulness votes was highly skewed, with many reviews receiving no votes at all (52.8%) or only one vote (21.5%). The distribution of useful votes is illustrated in Fig. 1a.

Expressions of doubt. Reviews were coded using the Linguistic Inquiry Word Count (LIWC) software (Pennebaker et al., 2015). The LIWC software uses a closed-vocabulary approach, meaning that texts were scored based on how often they contained terms from pre-defined dictionaries. LIWC dictionary scores range from 0 to 100, indicating the relative frequency at which terms from a dictionary appeared in a text.

Two LIWC dictionaries were related to the concept of doubt: tentativeness (example words: "maybe" and "perhaps"; $M = 2.05$; $SD = 2.24$) and certainty (example words: "always" and "never"; $M = 1.96$; $SD = 2.05$). Although the average relative frequencies of tentativeness and certainty were low, the majority of reviews (76.1%) contained at least one word from either dictionary. We created an overall measure of doubt, averaging tentativeness and (reverse-scored) certainty ($M = 50.14$, $SD = 1.50$, see Fig. 1b). We also conducted analyses using each of the two dictionaries separately. In other words, we also tested whether reviews containing more tentative language were more likely to receive useful votes, and whether reviews containing more certain language were less likely to receive useful votes. These additional analyses allowed us to test whether there were asymmetric effects of expressing doubt versus confidence.

Validation of the doubt dictionary. We conducted a study to validate this measure of doubt. We collected data from human raters to validate that reviews with higher doubt scores were actually seen as more doubtful than reviews with lower doubt scores. We randomly selected a set of 40 reviews that were between 50 and 100 words long (most reviews fall within this range). There were 10 low-doubt (–2SD) 1-star reviews; 10 high-doubt (+2SD) 1-star reviews; 10 low-doubt (–2SD) 5-star reviews; and 10 high-doubt (+2SD) 5-star reviews.

We recruited 200 participants from Prolific Academic (Prolific, 2019) to evaluate the authors of each review (three did not complete the study, final N = 197). The study took about 12 min and participants were paid £2 each for their time. Participants read each review and were asked to evaluate the confidence of the author on a 10-point scale, from 0 (not at all confident) to 10 (extremely confident). The 40 reviews were presented in a randomized order.

Our main analysis was a multilevel linear regression with perceived confidence as a dependent variable. The model included doubt (−1 = low-doubt scores; +1 = high-doubt scores), review positivity (−1 = 1-star reviews; +1 = 5-star reviews), and a doubt-by-positivity interaction term. The model also included random-intercepts for each participant and each review. Critically, high-doubt reviews were seen as less confident than low-doubt reviews, $b = -0.35$, $SE = 0.94$, $p < .001$. Positive reviews were also seen as more confident than negative reviews, $b = 0.65$, $SE = 0.94$, $p < .001$. There was no significant interaction between doubtful language and review positivity, $b = -0.050$, $SE = 0.94$, $p = .59$.

3.1.2. Additional measures

Psychological measures. We also conducted analyses including linguistic cues potentially associated with both doubt and perceptions of trustworthiness. We analyzed four LIWC summary scores derived from previous research linking each variable to specific outcomes: The analytical thinking variable was correlated with academic success (Pennebaker, Chung, Frazee, Lavergne, & Beaver, 2014); clout was correlated with social status (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2014); authenticity was correlated with deception (Newman, Pennebaker, Berry, & Richards, 2003); and emotional tone was correlated with the expression of positive (or negative) emotions (Cohn,

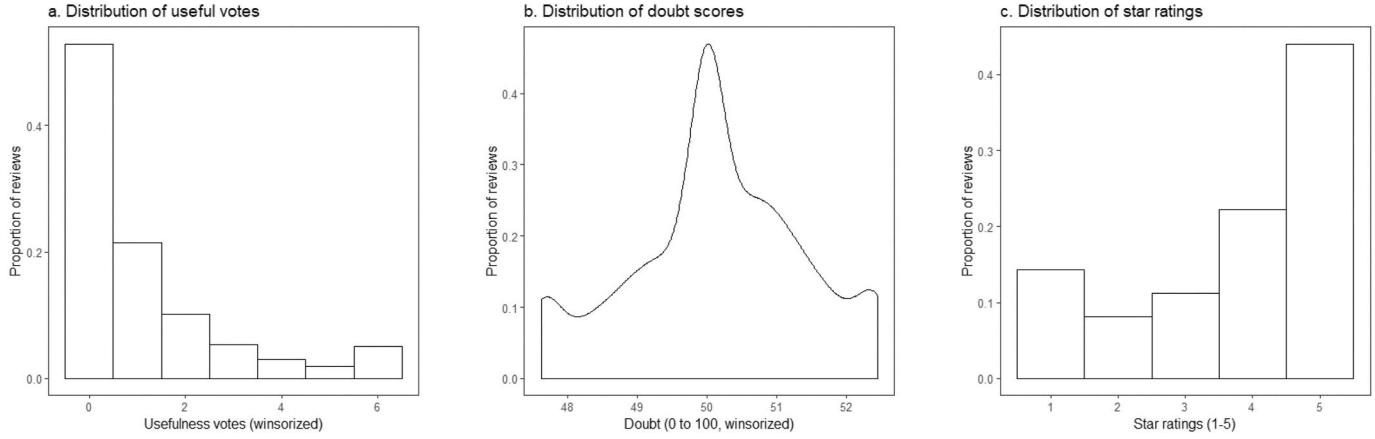

Fig. 1. The distributions of usefulness votes (a), doubt scores (b), and star ratings (c) in the Yelp dataset (Study 1). Our visualizations winsorized the top 5% of usefulness votes, and the top and bottom 5% of doubt scores (Signorell, Aho, Alfons, Anderegg, & Aragon, 2016). We used winsorization because there were a small number of observations with extreme which substantially reduced the figure's readability. Importantly, our primary analyses were conducted using the unwinsorized variables.

Mehl, & Pennebaker, 2004). We tested whether the effects of doubt were robust while controlling for these four summary variables.

Indicators of length and linguistic complexity. We also estimated models to control for three variables related to length and linguistic complexity: review length ($M = 111.47$ words, $SD = 105.42$), words per sentence ($M = 13.69$, $SD = 7.97$), and the relative frequency of words with more than six letters ($M = 15.70$, $SD = 6.01$). We included these measures to test if the effect of doubt remained robust when controlling for the perceived effort of reviewers. Doubtful reviewers may write longer and more complex reviews, and people may consider more elaborate reviews as being of higher quality (Kruger, Wirtz, Van Boven, & Altermatt, 2004).

Star ratings. Reviewers assigned star-ratings in their reviews, which ranged from 1-star (14.3%) to 5-stars (44.0%): $M = 3.73$ stars, $SD = 1.44$. The distribution of star ratings is illustrated in Fig. 1c. As in previous studies, positive reviews were much more common than negative reviews (De Langhe et al., 2015; Kupor & Tormala, 2018).

Analysis strategy. We conducted our analyses using functions from R and Stata. Given the large sample size, confidence intervals and measures of significance were generally uninformative; all statistical tests were significant at $p < .001$. Therefore, we omitted this information and focused on effect sizes.

Our central dependent variable (the number of useful votes per review) was an over-dispersed count variable (Fig. 1a); therefore, in our primary analyses we estimated a series of negative binomial regressions (Gardner, Mulvey, & Shaw, 1995). In this data set, single authors sometimes provided multiple reviews (1.5 million users wrote 5.9 million reviews). To account for this non-independence of observations, we estimated robust standard errors clustered at the level of author.

Additionally, we conducted supplemental analyses to examine the consistency of our results: First, we examined whether results were consistent across the ten largest cities in the dataset, with the finding that the effects of doubt on trust were generally homogenous. Second, we used two alternative analysis approaches: 1) we winsorized our dependent variable and estimated standard Ordinary Least Squares regressions, and 2) we dichotomized our dependent variable (0 = no useful votes; 1 = one-or-more useful votes) and estimated a series of logistic regressions. The results from these approaches were consistent with our negative binomial regressions (see Supplemental Materials). Finally, we also conducted additional analyses where we examined the effects of interrogatives (words used in questions) and reviewer experience on trust.

3.1.3. Results

Descriptive statistics and zero-order correlations of our primary variables of interest are reported in Table 1.

Expressions of doubt and perceived trustworthiness. Our first set of analyses focused on the relationship between expressions of doubt and the amount of useful votes received (our proxy for perceived trustworthiness). The results are reported in Table 2 and visualized in Fig. 2.

To begin, we estimated a negative binomial regression with useful votes as the dependent variable and doubt as the predictor (Table 2, Model 1). There was a positive relationship between doubtfulness and useful votes: $b = 0.11$.[3] In other words, a one-unit increase in doubt was associated with an 11% increase in the expected number of useful votes.

Next, we estimated a second regression testing the effect of doubt while controlling for the other psychological characteristics of reviews (Table 2, Model 2). We included LIWC measures of analytical thinking, clout, authenticity, and emotional tone. The positive relationship between doubt and positive votes remained significant: $b = 0.093$. By comparison, the other variables were only weakly associated with positive votes: b's ≤ 0.01 (though given the sample size, these effects were also significant). Comparing Model 1 and Model 2, adding these psychological variables reduced the effect of doubt by a factor of 15.7%.

In our third model, we controlled for the variables included in Model 2, as well as measures of review length, words per sentence, and the frequency of 6+ letter words (Table 2, Model 3). The effect of doubt remained positive, $b = 0.033$. Comparing Model 2 and Model 3, adding these three variables as controls reduced the effect of doubt on review usefulness by a factor of 64.3%. To gauge the robustness of our results, we also estimated models including quadratic effects of our covariates (e.g., the quadratic effect of emotional tone), and tested the effects of doubt using Ordinary Least Squares and logistic regressions. These approaches yielded similar results, which are reported in the supplemental materials.

Differentiating the effects of certainty and tentativeness. Our measure of doubt was based on the combination of two LIWC dictionaries, tentativeness and (lack of) certainty. This raises the possibility that results were driven primarily by one of the two sub-dictionaries (for example, there may be a negative relationship between certainty and usefulness, but no positive relationship between tentativeness and usefulness). Therefore, we tested the separate effects of certainty and tentativeness. Expressions of tentativeness were *positively* associated with usefulness ($b = 0.048$) and expressions of certainty were *negatively* associated with usefulness ($b = -0.066$). In other words, doubt is associated with increased usefulness and confidence is associated with decreased usefulness.

Assessing the effect size of doubt. Aside from our measure of doubt, the LIWC software also included measures of 38 additional psychological and linguistic categories. These categories, for example, included references to emotions (e.g., anxiety, anger, and sadness), psychological drives (e.g., affiliation, power, and reward), and social concerns (e.g., friends and family). Example words for each category are included in the Supplemental Materials. To gauge the relative importance of doubt, we compared its unstandardized effect size (estimated from Models 1, 2, and 3) with the absolute effect sizes of these 38 other psychological variables. In other words, we estimated a series of negative binomial regressions with useful votes as the dependent variable. The distribution of effect sizes is illustrated in Fig. 3. Among zero-order effect sizes (Model 1), the effect size for doubt was larger than 33 of the 38 effects (86.8%). When estimating effect sizes while including all covariates (Model 3), the effect size for doubt was larger than 28 of the 38 effects (73.7%).

Another way of assessing the effect size of doubt is to compare its effect to a relevant non-linguistic variable. To this end, we estimated the effect of number of years since the publication of each review, $M = 3.08$ years, $SD = 2.37$.[4] Usefulness votes should increase with time since publication, as the passage of time allows more readers to read and evaluate reviews. Not surprisingly, there was a positive relationship between years since publication and the number of useful votes, $b = 0.127$; each year after publication is associated with a 13.5% increase in the number of expected useful votes. To put our doubt effect in context, increasing the relative frequency of doubt in a review by one unit is equivalent to keeping a review online for a period of approximately 9.6 months (or 2.97 additional months, if we use the most conservative estimate of doubt's effect).

Does review positivity moderate the effect of doubt? We also examined whether the effects of doubt were different for positive versus negative user reviews. To investigate this question, we estimated a regression predicting positive votes with the following variables entered as predictors: doubt (centered), number-of-stars (centered), and a doubt by number-of-stars interaction term. The main effect of doubt was

[3] b indicates the unstandardized regression coefficient. LIWC scores ranged from 0 to 100. Hence, the effects of different LIWC variables can be directly compared.

[4] We calculated the days between each review's publication date and the latest publication date in the dataset, then divided this number by 365.

Table 1
Means, standard deviations, and correlations (Study 1).

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. Useful votes	1.37	3.71										
2. Doubt	50.14	1.51	.04									
3. Analytic	55.10	25.40	.03	.05								
4. Clout	52.49	25.87	-.00	-.10	.11							
5. Authentic	47.67	30.86	.03	.03	-.08	-.41						
6. Tone	75.93	31.32	-.08	-.08	.04	.10	-.18					
7. Word count	111.4	105.1	.26	.14	.04	-.00	.09	-.17				
8. Words per sentence	13.69	7.97	.07	.09	.02	-.02	.10	-.14	.26			
9. Six + Letter words	15.70	6.01	-.01	-.12	.14	.10	-.21	.08	-.10	-.10		
10. Stars	3.74	1.45	-.09	-.16	.06	.16	-.15	.62	-.20	-.13	.13	
11. Years since pub	3.09	2.37	.11	.09	.07	-.07	.03	.01	.14	.04	-.07	-.03

Table 2
The positive association between doubtful language and usefulness votes (Study 1).

	Model 1		Model 2		Model 3	
	b	SE	b	SE	b	SE
Constant	−5.11	0.12	−4.25	0.12	−2.11	0.079
Doubt	0.108	0.0025	0.093	0.0024	0.034	0.0015
Analytical thinking			0.0038	0.00021	0.0017	0.00013
Clout			0.0015	0.00015	0.00037	0.00010
Authenticity			0.0021	0.00014	0.00086	0.000094
Emotional tone			−0.0070	0.00011	−0.0047	0.000084
Word count					0.0049	0.000059
Words-per-sentence					0.0029	0.00047
Six-letter-words					0.0039	0.00047

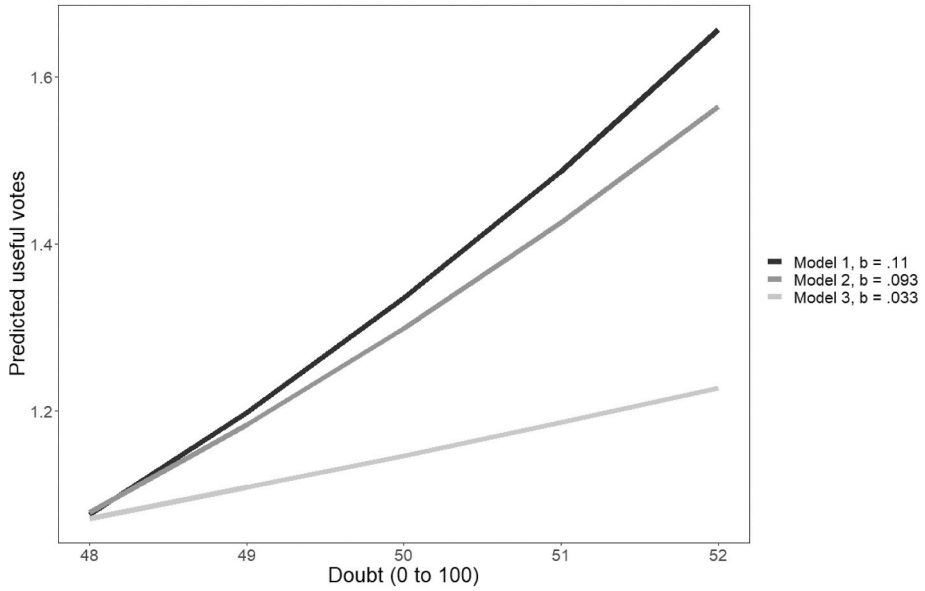

Fig. 2. The predicted relationship between doubt and usefulness votes. Model 1 shows the zero-order relationship between doubt and usefulness; Model 2 adds LIWC measures of clout, authenticity, emotional tone, and analytical thinking; Model 3 includes these psychological control variables, as well as measures of review length, words per sentence, and the frequency of 6+ letter words (Study 1).

positive ($b = 0.080$) and the main effect of stars was negative ($b = -0.149$); doubtful reviews and negative reviews received more useful votes. Critically, there was also a positive doubt by number-of-stars interaction ($b = 0.027$). Fig. 4 illustrates the predicted interaction between review valence and doubt; the effect of doubt was significantly stronger for positive (vs. negative) reviews.

Additional analyses. Our supplemental materials also include two sets of additional analyses: First, we examined the effects of interrogatives, words used to ask questions (such as "how", "what", and "when"). We used interrogatives as an alternative measure of reviewer doubt. Presumably, reviews containing questions are perceived as more doubtful. Consistent with our main results, reviews containing interrogatives received more useful votes, and this effect was robust when also controlling for doubt.

Second, we examined whether reviewer experience moderated the effect of doubt on perceive usefulness. Our measure of experience was the total number of reviews written by each author in the dataset. Reviews from experienced authors were more likely to receive useful votes, and author experience moderated the effect of doubt on usefulness. There was a negative interaction, such that the positive effect of doubt

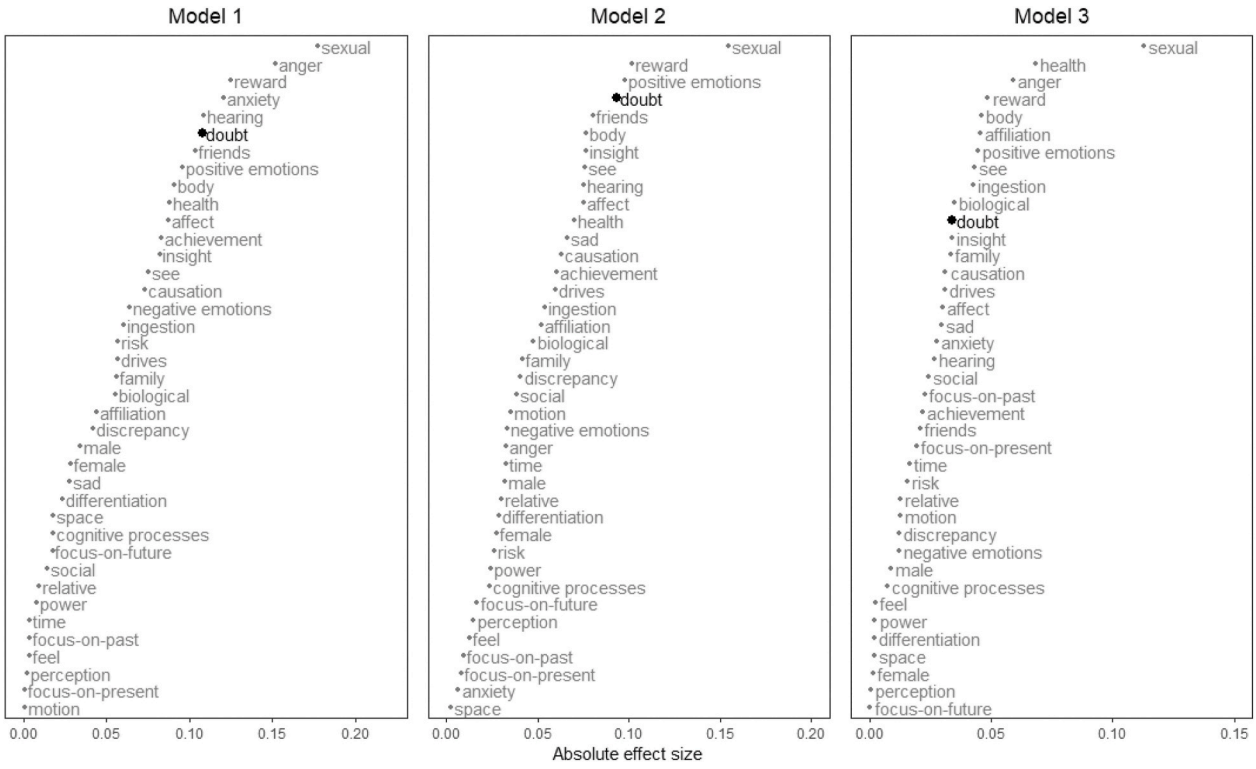

Fig. 3. The effect of doubt (black) was compared with 38 other psychological dictionaries (gray) measured using LIWC (Study 1). Model 1 analyses included no covariates; Model 2 analyses included the four LIWC summary scores as covariates; and Model 3 analyses included both summary scores and linguistic measures as covariates.

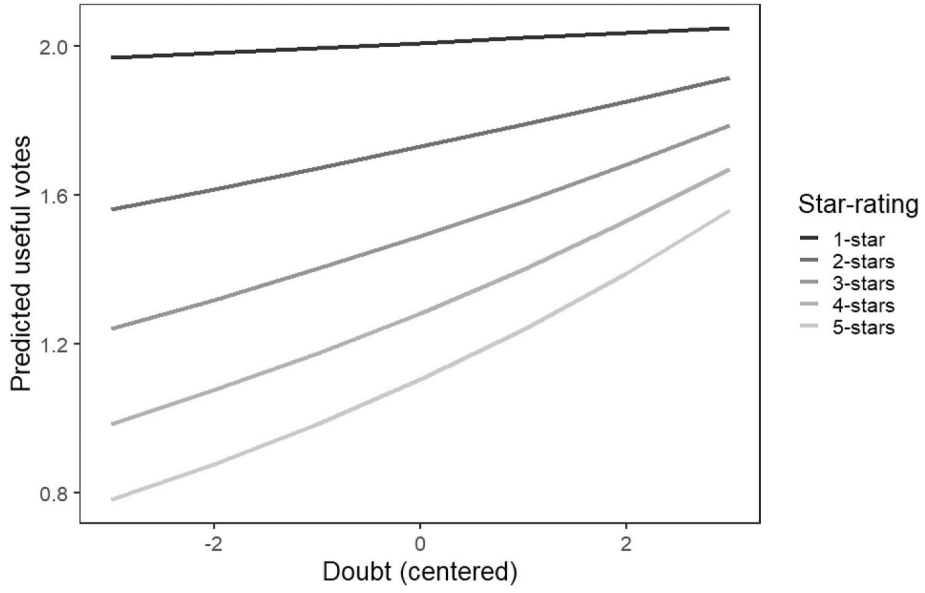

Fig. 4. The interactive effects of doubt and valence on review usefulness (Study 1).

was strongest for inexperienced authors. These results are consistent with the broader conclusion that doubt has positive effects when readers are most concerned about the trustworthiness of sources (in this case, when reading a review from a first-time author).

Summary. Expressions of doubt were associated with increased trust in online reviews. This effect was robust when controlling for other psychological variables (such as analytical thinking style and emotional tone) and measures of length and linguistic complexity. Moreover, the effects of doubt were stronger for positive (vs. negative) reviews.

3.2. Study 2

We conducted an experiment to test the robustness of Study 1's two main results: the positive effect of doubt on review trustworthiness and the interaction between review valence and doubt (i.e., the finding that the positive effect of doubt was strongest for positive reviews). Participants read and evaluated the authenticity (i.e., the likelihood that the review was written by a real consumer) of a series of positive and negative reviews; critically, we systematically manipulated whether

these reviews contained doubtful or confident conclusions (expressed through doubtful or confident language). The study was pre-registered at https://aspredicted.org/blind.php?x=fb7nf8 (for more information about the use of pre-registration, see van't Veer & Giner-Sorolla, 2016) and the study was approved by the Ethics Review Board at the host institution. Materials and data are available at https://osf.io/w5dy7/?view_only=420062c267084a5fbf1c16146b8da3df.

3.2.1. Method

Materials. Participants read 16 Yelp reviews for restaurants in the California Bay Area. Reviews underwent minor copy editing for grammar and spelling. Eight of the reviews were positive (e.g., "Very yummy food, we enjoyed every bite.") and eight were negative (e.g., "Extremely disappointed with this restaurant."). We edited the final sentence of each review to create doubtful and confident versions. Doubtful example: "I would probably recommend this place"; Confident example: "Without a doubt, I would recommend this place." The full text of each review is included in the supplemental materials.

Pre-test. We conducted a pre-test to verify that the doubtful versions of the reviews were perceived by readers as being more doubtful than the confident versions. We recruited 100 participants from Prolific Academic (51 Men, 47 Women, and 2 gender non-binary participants; M_{age} = 31.27 years; SD_{age} = 11.19) to read and assess the confidence of the author of each review. Each participant read four doubtful-positive reviews, four confident-positive reviews, four doubtful-negative reviews, and four confident-negative reviews. We created two variants of the survey, randomizing which reviews were presented as confident versus doubtful. Participants were asked to rate the confidence of the author of each review on a scale from 0 (Not confident at all) to 10 (Extremely confident).

We estimated a multilevel linear model with confidence ratings as the dependent variable. The model included doubt (−1 = confident; +1 = doubt), review valence (−1 = negative, +1 = positive), and a doubt-by-valence interaction term; the model also included random intercepts for each participant and each review. There were significant main effects of doubt and valence: doubtful reviews were seen as less confident than confident reviews ($b = -0.90$, $SE = 0.046$, $p < .001$) and positive reviews were seen as more confident than negative reviews ($b = 0.36$, $SE = 0.11$, $p = .004$). There was no significant interaction between review valence and doubt, $b = -0.029$, $SE = 0.046$, $p = .52$.

Procedure. Each participant evaluated four doubtful-positive reviews, four confident-positive reviews, four doubtful-negative reviews; and four confident-negative reviews. Participants were randomly assigned to one of two versions of the experiment: these versions only differed in terms of which reviews were presented as confident vs. doubtful. In other words, we used a mixed design with three factors: valence (positive vs. negative) was manipulated within-subjects; doubt (vs. confidence) was manipulated within subjects; and survey version (which reviews were presented as doubtful vs. confident) was manipulated between subjects. The order in which the reviews were presented was randomized. Participants rated the authenticity of each review on a 10-point scale: "How likely is it that this review is authentic (written by an actual consumer)?" where 0 = "Definitely fake" and 10 = "Definitely authentic."

Participants. Our initial sample size was based on the number of participants needed to detect a small effect ($d = 0.20$) using a paired-sample t-test with 80% power and $\alpha = 0.05$: Minimum $N = 199$; Total $N = 220$. However, our first analyses yielded inconclusive results for the main effect of doubt ($p = .07$) and for the hypothesized doubt-by-positivity interaction ($p = .14$). We decided to triple our total sample size, resulting in a final $N = 660$. This second wave of data collection inflated our Type-I error rate; therefore, we adopted a more conservative $\alpha = 0.025$ (Sagarin, Ambler, & Lee, 2014). U.S. American participants were recruited through Prolific Academic. The average age was 36.45; $SD_{age} = 12.07$; there were 326 men, 328 women, and 6 participants who provided other gender responses. Participants were paid £1.20 each for their time.

3.2.2. Results

We estimated a multilevel linear model with perceived authenticity as the dependent variable. Doubt (−1 = confident; +1 = doubtful), review valence (−1 = negative; +1 = positive), and a doubt-by-positivity interaction term were entered as predictors. The model also included random-intercepts for each participant and each review. Consistent with our first hypothesis, doubtful reviews were seen as more authentic than confident reviews: $b = 0.12$, $SE = 0.022$, $p < .001$. Positive reviews were also seen as less authentic than negative reviews, though this effect was not significant at our adjusted $\alpha = 0.025$: $b = -0.24$, $SE = 0.11$, $p = .051$. Consistent with our second hypothesis, we detected a significant doubt-by-valence interaction: $b = 0.079$, $SE = 0.023$, $p < .001$. The effects of doubt and review valence on perceived authenticity are illustrated in Fig. 5.

To better understand the interaction between doubt and valence, we estimated the simple effects of doubt for positive and negative reviews. Positive reviews were seen as more authentic when they contained a doubtful (vs. confident) conclusion, $b = 0.20$, $SE = 0.03$, $p < .001$. In contrast, doubt had a positive, but not significant, effect on the perceived authenticity of negative reviews, $b = 0.044$, $SE = 0.027$, $p = .10$.

Summary. Replicating the key findings from our Yelp analyses, doubtful reviews were seen as more authentic than confident reviews. Moreover, the effect of doubt interacted with review valence: Doubtful positive reviews were seen as more authentic than confident positive reviews. However, expressions of doubt had no significant effect on the perceived authenticity of negative reviews.

3.3. General discussion

Do expressions of doubt help (or hurt) trust in online advice? Our analyses of Yelp data and a controlled experiment showed that expressions of doubt are associated with increased, rather than decreased, trust in online reviews. Reviews with tentative language ("maybe" and "perhaps") were more likely to be seen as useful and authentic compared to reviews which contained confident language ("always" and "never"). Additionally, the effect of doubt was stronger for positive, compared to negative, reviews.

3.4. The interpersonal consequences of expressing doubt

Our results add to a recent literature which suggests that, under certain circumstances, expressions of confidence can undermine trust in advice (Tenney, Meikle, Hunsaker, Moore, & Anderson, 2018). We find that doubt is seen as trustworthy (and confidence is seen as untrustworthy) in the domain of online user reviews, and that the effects of doubtful language on trust are stronger for positive (vs. negative) reviews. These results are in conflict with the predictions of the confidence

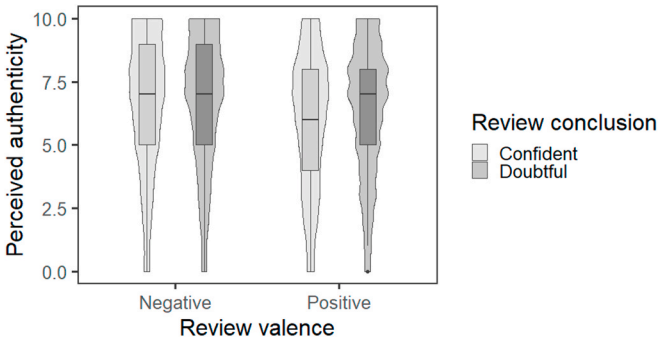

Fig. 5. The effects of doubt and review valence on perceived authenticity (Study 2).

heuristic, which suggests that confident reviews should be trusted more than doubtful reviews (Price & Stone, 2004).

Why are doubtful consumer reviews trusted more than confident reviews? We briefly consider two possible explanations: One possibility is that readers assume that doubtful reviewers spent more time generating their opinions, and consider this greater time investment as an indicator of reviewer effort. In line with this reasoning, Efendić, Van de Calseyde, and Evans (2020) found that people are more likely to trust slowly (vs. quickly) generated predictions from human forecasters, in part because of the assumption that slow forecasters put more effort into their predictions. Similarly, people may believe that doubtful authors put more effort into their reviews, as slow responses are often assumed to be more doubtful (Evans & Van De Calseyde, 2017). There may be justification for this belief, as the use of doubtful language in reviews is positively correlated with review length and sentence complexity. While this effect is plausible, it does not readily explain why the positive effects of doubt are stronger for positive (vs. negative) reviews.

Alternatively, readers may perceive doubt as a costly signal of honesty that can mitigate concerns about reviewers' ulterior motives and conflicts-of-interest (Polnaszek & Stephens, 2014). From an author's perspective, admitting doubts means incurring a potential cost, as doubt could be seen as a sign of incompetence. Importantly, reviewers with ulterior motives (e.g., owners who write fake positive reviews to boost business) should be less willing to incur this potential cost. Therefore, honest reviewers should be more likely to use doubtful language. This reasoning would explain why doubt has particularly strong positive effects for positive reviews, as users are concerned about the honesty of positive online reviews (Luca & Zervas, 2016) and expressions of doubt might be seen as a cue to honesty and mitigate this concern. Note that the conflict-of-interest account suggests that doubt is likely to enhance trustworthiness (and persuasion) in low-trust online environments, where readers are concerned about the presence of bots, scammers, or other malicious agents.

3.5. Practical implications

The present research has implications for both review authors and platforms that publish online reviews, such as Yelp and TripAdvisor. First, our results suggest that online advice-givers may be able to increase their persuasiveness by not appearing overly confident, as readers may discount the opinions of overly certain advice. This insight is in stark contrast to previous recommendations, which emphasized the positive consequences of giving confident advice (Bonaccio & Dalal, 2006; Kahneman, 2011). Our results are also of interest to platforms that publish online reviews: Knowing that users prefer doubtful reviews, platforms may be able to increase user engagement by making doubtful reviews more accessible to users, either by highlighting them or prioritizing them in users' search results. Alternatively, it may be useful to use (lack of) doubt in text as a cue to flag potential fake reviews; however, further work would be needed to verify the feasibility of this approach.

3.6. Limitations

Our central finding, the relationship between expressions of doubt in a review and positive votes received, appears to be relatively small. However, there are also good reasons to believe our findings have practical significance: In Study 1, the effect of doubt was relatively large when compared to other linguistic variables; in fact, the zero-order effect of doubt was larger than 86.8% of the effects of linguistic variables. Furthermore, the effect of doubt can also be evaluated by comparing it to other relevant factors that influence reactions to reviews. Increasing a review's doubt score by one is equivalent to leaving a review posted for a period of 9.6 additional months. If we adopt the most conservative effect size estimate (from Model 3, which included all covariates), the effect of doubt is larger than 73.6% of linguistic variables and increasing the doubt score by one is equivalent to leaving a review online for a period of 2.97 additional months. Though the effect of doubt is small, it is robust, theoretically interesting, and practically important.

Analyses of large scale datasets raise concerns about the possibility of spurious findings and analytical degrees of freedom (Orben & Przybylski, 2019), and these concerns rightfully apply to our first study. We note two salient limitations: First, we relied on closed-dictionary methods (i.e., LIWC) to measure doubt. This approach makes it difficult to know about the contexts in which authors expressed their doubts. For example, our analyses did not differentiate between self-focused doubts ("I'm not sure if I can recommend this place") and business-related doubts ("I doubt if they can make a decent cup of coffee"). Second, we cannot rule out the role of potential confounding variables. Note that the effect of doubt was substantially attenuated (by 64.3%) when we controlled for review length and linguistic complexity (i.e., Model 3 in our analyses). This raises the possibility that the effects of doubt may be related to overarching differences in review quality or reviewer effort, though our second study (that manipulated the level of doubt in a controlled experiment) gives us confidence that doubt does directly influences perceptions of reviewer trustworthiness. Future research should examine the factors, such as cognitive effort and financial conflicts-of-interest, that encourage authors to express doubts.

Finally, it is important to note that our studies focused on the characteristics of review authors, and did not examine how user characteristics may also influence trust in online reviews. For example, future research should consider how prior experience with online reviews platforms changes the way that people perceive and use online reviews.

4. Conclusion

How do people evaluate the veracity of online advice? This question is important to our understanding of consumer decision making, and more generally, online behavior. In contrast to the idea that confidence increases trust in online advice, we find evidence that doubt (rather than confidence) is associated with increased trust. People are more likely to consider doubtful reviews as useful, and the beneficial effects of doubt are stronger for positive (vs. negative) reviews. The present research adds to our understanding of the cues that influence online advice-taking behavior, and the interpersonal consequences of expressing doubt.

Authorship credit

Anthony M Evans: Conceptualization, Methodology, Investigation, Formal Analysis, Visualization, Writing - Original Draft. **Olga Stavrova**: Conceptualization, Methodology, Writing - Review & Editing. **Hannes Rosenbusch**: Conceptualization, Methodology, Writing - Review & Editing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chb.2020.106556.

References

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101*(2), 127–151.

Butler, P. (2018). *Convert the Yelp academic dataset from JSON to CSV files with pandas*. Retrieved from https://gist.github.com/paulgb/5265767/.

van de Calseyde, P. P., Zeelenberg, M., & Evers, E. R. (2018). The impact of doubt on the experience of regret. *Organizational Behavior and Human Decision Processes, 149*, 97–110.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science, 15*(10), 687–693.

De Langhe, B., Fernbach, P. M., & Lichtenstein, D. R. (2015). Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research, 42*(6), 817–833.
Efendić, E., Van de Calseyde, P. P., & Evans, A. M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes, 157*, 103–114.
Evans, A. M., & Krueger, J. I. (2016). Bounded prospection in dilemmas of trust and reciprocity. *Review of General Psychology, 20*(1), 17.
Evans, A. M., & Van De Calseyde, P. P. (2017). The effects of observed decision time on expectations of extremity and cooperation. *Journal of Experimental Social Psychology, 68*, 50–59.
Filieri, R., Alguezaui, S., & McLeay, F. (2015). Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth. *Tourism Management, 51*, 174–185.
Filieri, R., Hofacker, C. F., & Alguezaui, S. (2018). What makes information in online consumer reviews diagnostic over time? The role of review relevancy, factuality, currency, source credibility and ranking score. *Computers in Human Behavior, 80*, 122–131.
Gaertig, C., & Simmons, J. P. (2018). Do people inherently dislike uncertain advice? *Psychological Science, 29*(4), 504–520.
Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin, 118*(3), 392.
Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology, 33*(2), 125–143.
Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
Kim, J., & Gupta, P. (2012). Emotional expressions in online user reviews: How they influence consumers' product evaluations. *Journal of Business Research, 65*(7), 985–992.
Kruger, J., Wirtz, D., Van Boven, L., & Altermatt, T. W. (2004). The effort heuristic. *Journal of Experimental Social Psychology, 40*(1), 91–98.
Kupor, D., & Tormala, Z. (2018). When moderation fosters persuasion: The persuasive power of deviatory reviews. *Journal of Consumer Research, 45*(3), 490–510.
Lee, M., Jeong, M., & Lee, J. (2017). Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website. *International Journal of Contemporary Hospitality Management, 29*(2), 762–783.
Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science, 62*(12), 3412–3427.
Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. In *Paper presented at the proceedings of the 21st international conference on world wide web*.
Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*(5), 665–675.
Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour, 1*.
Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015* (Retrieved from).
Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PloS One, 9*(12), Article e115844.
Pew Research Center. (2016). *Online shopping and e-commerce*. Retrieved from https://www.pewinternet.org/2016/12/19/online-reviews/.
Polnaszek, T. J., & Stephens, D. W. (2014). Why not lie? Costs enforce honesty in an experimental signalling game. *Proceedings of the Royal Society B: Biological Sciences, 281*(1774), 20132457.
Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making, 17*(1), 39–57.
Prolific. (2019). Prolific Academic.
Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science, 9*(3), 293–304.
Sah, S., Moore, D. A., & MacCoun, R. J. (2013). Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes, 121*(2), 246–255.
Sen, S., & Lerman, D. (2007). Why are you telling me this? An examination into negative consumer reviews on the web. *Journal of Interactive Marketing, 21*(4), 76–94.
Signorell, A., Aho, K., Alfons, A., Anderegg, N., & Aragon, T. (2016). *DescTools: Tools for descriptive statistics. R package version 0.99. 18*. Vienna, Austria: R Found. Stat. Comput.
Stavrova, O., & Evans, A. M. (2019). Examining the trade-off between confidence and optimism in future forecasts. *Journal of Behavioral Decision Making, 32*(1), 3–14.
Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science, 18*(1), 46–50.
Tenney, E. R., Meikle, N. L., Hunsaker, D., Moore, D. A., & Anderson, C. (2018). Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. *Journal of Personality and Social Psychology, 116*(3), 396–415.
Thomas, J. P., & McFadyen, R. G. (1995). The confidence heuristic: A game-theoretic analysis. *Journal of Economic Psychology, 16*(1), 97–113.
Van de Calseyde, P. P., Keren, G., & Zeelenberg, M. (2014). Decision time as information in judgment and choice. *Organizational Behavior and Human Decision Processes, 125*(2), 113–122.
van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—a discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2–12.
Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83*(2), 260–281.
Yelp. (2017). *Yelp open dataset*. Retrieved from https://www.yelp.com/dataset.
Yin, D., Bond, S. D., & Zhang, H. (2014). Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quarterly, 38*(2), 539–560.