AMA AMERICAN MARKETING ASSOCIATION

Article

Journal of Marketing Research
1-19

DOI: 10.1177/00222437221134802
journals.sagepub.com/home/mrj
SAGE

Beyond Sentiment: The Value and Measurement of Consumer Certainty in Language

Matthew D. Rocklage, Sharlene He, Derek D. Rucker, and Loran F. Nordgren

Abstract

Sentiment analysis has fundamentally changed marketers' ability to assess consumer opinion. Indeed, the measurement of attitudes via natural language has influenced how marketing is practiced on a day-to-day basis. Yet recent findings suggest that sentiment analysis's current emphasis on measuring valence (i.e., positivity or negativity) can produce incomplete, inaccurate, and even misleading insights. Conceptually, the current work challenges sentiment analysis to move beyond valence. The authors identify the certainty or confidence of consumers' sentiment as a particularly potent facet to assess. Empirically, they develop a new computational measure of certainty in language—the Certainty Lexicon—and validate its use with sentiment analysis. To construct and validate this measure, the authors use text from 11.6 million people who generated billions of words, millions of online reviews, and hundreds of thousands of entries in an online prediction market. Across social media data sets, in-lab experiments, and online reviews, the authors find that the Certainty Lexicon is more comprehensive, generalizable, and accurate in its measurement compared with other tools. The authors also demonstrate the value of measuring sentiment certainty for marketers: certainty predicted the real-world success of commercials where traditional sentiment analysis did not. The Certainty Lexicon is available at www.CertaintyLexicon.com.

Keywords

certainty, confidence, sentiment analysis, word of mouth, attitudes, language

Online supplement: https://doi.org/10.1177/00222437221134802

Sentiment analysis enables marketers to discern consumer opinion on a scale never before possible. By quantifying the positivity of consumers' language, companies can measure consumers' likes and dislikes (i.e., their attitudes) toward anything from new products, to advertisements, to the company itself. Sentiment analysis appears to hold the promise of understanding consumer attitudes and providing insight into their future behavior, all in real time.

For example, Bloomberg acquires market insight by converting "a news article about a company into a number that expresses positive … and negative sentiment" (Mighty Guides 2016, p. 5). Microsoft uses sentiment analysis "to maintain, collaborate, and distribute results that help our senior leadership teams make better decisions" (Dancshazy 2021). Similarly, academics use sentiment analysis to understand and predict consumer behavior (Berger 2014; Humphreys and Wang 2018; Moore and Lafreniere 2020).

The valence (positivity or negativity) of people's language is the most researched aspect of word of mouth (Chen and Yuan 2020). Moreover, an audit of popular sentiment analysis providers reveals that their tools focus almost exclusively on quantifying the valence of consumers' attitudes (see Web Appendix A). Yet, despite its importance, we argue that sentiment analysis's focus on valence is narrow and fundamentally incomplete. Indeed, valence by itself can be a poor, even misleading, predictor of behavior (Petty and Krosnick 1995; Tormala and Rucker 2018; Wicker 1969).

We propose and demonstrate that consumer sentiment is much richer than valence. As we detail, one particularly potent facet of consumers' sentiment is the certainty or confidence with which they hold that sentiment. The present research

Matthew D. Rocklage is Assistant Professor of Marketing, Northeastern University, USA (email: m.rocklage@northeastern.edu). Sharlene He is Assistant Professor of Marketing, Concordia University, Canada (email: sharlene.he@concordia.ca). Derek D. Rucker is Sandy and Morton Goldman Professor of Entrepreneurial Studies in Marketing, Northwestern University, USA (email: d-rucker@kellogg.northwestern.edu). Loran F. Nordgren is Professor of Management and Organizations, Northwestern University, USA (email: l-nordgren@kellogg.northwestern.edu).

aims to push sentiment analysis beyond valence. We develop a validated computational linguistic tool to measure the degree of certainty in language: the Certainty Lexicon (CL). To do so, we collect and analyze billions of words, millions of online reviews, and hundreds of thousands of entries in an online prediction market. Using these sources, we construct a comprehensive measure of certainty in language. We validate this tool as a measure of sentiment certainty specifically, but it can also be used to measure an individual's certainty for any belief or thought they hold.

To overview, we first discuss sentiment analysis and the shortcomings of focusing solely on valence. Then, drawing on attitudes research, we argue for the value of incorporating certainty into sentiment analysis. We also discuss how the CL addresses the shortcomings of existing tools. Finally, we present the construction of the CL, validate its ability to capture certainty in language, and demonstrate its predictive utility.

Sentiment Analysis

Researchers have used automated text analysis for 60 years (e.g., Stone et al. 1962). However, its use for sentiment analysis—the large scale measurement of attitudes from natural language—was almost impossible until recently. Advances in computing power and the availability of vast repositories of digitized text has enabled sentiment analysis to grow rapidly. In the last 10 years, online searches for sentiment analysis have increased over 1,400% (Google Trends 2021), academic publications on the subject have increased 2,500% (Web of Science 2021), and books that discuss the topic have increased 3,000% (Google Books Ngrams 2021).

Despite the explosion of sentiment analysis, marketers' tools continue to focus nearly exclusively on the simplest facet of consumer sentiment: whether it is positive or negative (i.e., its valence) (Lexico 2021).[1] Indeed, much of the interest and growth in this area can be attributed to work aimed at incrementally increasing the accuracy of measuring valence (e.g., Qin, Hu, and Liu 2020). Marketers have even suggested that if they could only measure valence with 100% accuracy, they could unlock all the knowledge consumers have to share with them (Burn-Murdoch 2013; Kessler 2014). Yet might this continued focus on valence ultimately lead marketers to miss out on other critical information?

Academic researchers note that it can be difficult to predict behavior based on consumer sentiment, even when measured under ideal conditions (Babić Rosario et al. 2016; De Langhe, Fernbach, and Lichtenstein 2016; Holbrook and Addis 2007). Though this difficulty can be attributed to multiple reasons (Babić Rosario et al. 2016), it suggests that even in an ideal world where consumers directly report their positivity toward a brand or product, valence provides an incomplete and sometimes even uninformative picture of consumer opinion.

Consumer Certainty

To better understand consumer sentiment, we argue that the certainty with which consumers hold that sentiment is pivotal. Certainty is an individual's subjective sense of confidence or conviction (Petrocelli, Tormala, and Rucker 2007). Research from the attitudes literature shows that the more certain a consumer is about an attitude or belief they hold, the more likely that attitude or belief will drive behavior (see Tormala and Rucker 2018). For example, research indicates there is a stronger association between attitudes and behavioral intentions when attitudes are held with more certainty ($r = .89$) than less certainty ($r = .68$; Tormala and Petty [2002], Experiment 4; see also Franc [1999]). Similarly, thoughts held with more certainty are more predictive of people's reliance on those thoughts (Briñol, Petty, and Tormala 2004). Within the attitudes literature, a large amount of research suggests that attitudes held with greater certainty are also more likely to persist over time and resist change (Rucker, Petty, and Briñol 2008; Tormala 2016; Tormala and Petty 2002).

Certainty is also distinct from the valence of individuals' sentiment (i.e., whether it is positive or negative) and the extremity of that valence (i.e., how positive or negative the valence is) (Clarkson, Tormala, and Leone 2011; Petty and Krosnick 1995). Although more extreme valence is often associated with more certain attitudes, this association is moderate (e.g., $r \sim .50$; Krosnick et al. 1993). Even when attitudes do not differ in extremity, differences in certainty can be prevalent (see Rucker and Petty 2004; Tormala and Petty 2002). Moreover, extreme attitudes can be held with less certainty (Litt and Tormala 2010) and be less likely to persist across time (Rocklage and Luttrell 2021). Thus, certainty and valence extremity are distinct.

To illustrate, consider two customers who visit the same restaurant and give it a perfect five-star rating. Though they are equally positive toward the restaurant, one may be more certain about their attitude because many of their friends hold a similar attitude (Tormala and DeSensi 2009). Despite having the exact same attitude valence, the customer with greater certainty is more likely to revisit the restaurant and recommend it to others (e.g., Barden and Petty 2008). Differences in certainty can arise from factors such as the amount of social consensus or direct personal experience. More generally, certainty can arise from any factors impacting consumers' sense that the information underlying their attitude or belief is accurate, complete, relevant, legitimate, or important (Rucker et al. 2014).

Given that certainty is an important and prominent facet of attitudes, it is an ideal candidate for expanding the assessment of consumer sentiment. At present, sentiment analysis largely focuses on measuring valence, but ignores the certainty associated with that valence. Moreover, research suggests that

[1] In recent years, some companies have begun to introduce language tools that attempt to identify discrete consumer emotions (e.g., happiness, disgust). However, these tools are largely undeveloped, and the validity of these measures remains unclear (Seyeditabari, Tabari, and Zadrozny 2018).

because most consumer sentiment expressed online is positive, there is often a "positivity problem" for marketers (Rocklage, Rucker, and Nordgren 2021b). This glut of positivity presents a restricted range of valence, which can make it difficult to gain insights based on valence or valence extremity alone. The measurement of certainty may be particularly useful in these contexts, and certainty may be a stronger predictor of behavior than valence.

In addition, despite natural parallels between attitudes research and sentiment analysis, there is little to no interaction between these literatures. The attitudes literature has identified certainty as a critical dimension for using attitudes to predict behavior. Yet sentiment analysis, perhaps surprisingly, does not consider certainty. At the same time, attitudes researchers have done little to construct computational tools that leverage the insights from their basic research.

The current work focuses on sentiment analysis given its importance to marketers. However, the tool we introduce is a measure of certainty in language and, as such, is designed to be applicable for broader use. For example, researchers or practitioners may wish to assess the degree of consumers' certainty about the economy, voters' certainty about political events, investors' certainty about the markets, or management's certainty about their company's prospects. These are all contexts where using a comprehensive tool to measure certainty in language can offer important insights.

The Measurement of Certainty in Language

Although there is a small set of tools that aim to measure certainty in language, they have significant limitations. Existing methods (1) have received surprisingly little formal validation, have not be tested for use with sentiment analysis, and have not been integrated into the sentiment analysis literature, and (2) are constrained in critical ways that limit their utility. The two most prominent measures of certainty in language are from the Linguistic Inquiry and Word Count (LIWC; Pennebaker et al. 2015) and DICTION (Hart and Carroll 2015) software programs. Both programs provide measures for assessing properties of text such as its valence. Although they also contain measures related to certainty, these measures are limited in their validity, generalizability, and the methodology they use to quantify language.

To begin, both LIWC and DICTION have received relatively little empirical validation for measuring certainty and have not been validated to assess sentiment certainty specifically. For example, both tools were created based on the researchers' intuition for the kind of words that would signal an individual's certainty as opposed to a more formal or data-driven approach (Hart 1976; Pennebaker and Francis 1996). LIWC contains two measures of certainty called "certainty" and "tentativeness." However, the "certainty" measure has not been directly validated (Petrie, Booth, and Pennebaker 1998), and the "tentativeness" measure has been validated only within a sample of 35 college students who wrote on the single topic of their college experience (Pennebaker and Francis 1996). Similarly, DICTION's measure of certainty has not been directly validated (Hart 1976, 1984). Although it is possible that they can work as measures of sentiment certainty, their validity and generalizability are ambiguous.

These tools also have three key methodological limitations. First, they both rely on a word count approach. Consider how LIWC quantifies these two sentences: (1) "I've often disliked my experience with that brand" and (2) "I've sorta disliked my experience with that brand." The words "often" and "sorta" are both present in LIWC's "tentativeness" (uncertainty) word list. Under LIWC's word count approach, both sentences are therefore given a score of 12.50% (i.e., one out of eight words signals uncertainty). Thus, "often" and "sorta" are counted as signaling equal levels of uncertainty. Intuition suggests, however, that these words convey quite different levels of certainty. Similarly, DICTION gives these sentences the same score on certainty.[2] By simply counting the words in each sentence, word count approaches treat all words on a given word list as signaling identical certainty.

A second major limitation of word count approaches is that they can provide poorer measures in shorter pieces of text such as those found on Twitter and Facebook. This is because short pieces of text contain relatively little information and therefore often have only a single keyword related to certainty (Pennebaker et al. 2015). Given that word count approaches assume that all words in a given dictionary signal the same level of certainty, these measures can lead to a large skew in the data (observations with little variability), resulting in a great deal of noise and therefore uninformative or even misleading results (Garten et al. 2018; Rocklage and Rucker 2019; Sterling, Jost, and Bonneau 2020). This limitation is of particular relevance to marketers given their reliance on social media for understanding consumer sentiment (Schaefer 2015).

A third methodological limitation of prior tools is that they only analyze single words and cannot process phrases. For example, both LIWC and DICTION would treat the phrase "I'm not sure" as indicating high certainty because it includes the word "sure"; these approaches cannot recognize the key phrase "not sure." Similarly, they would treat "likely" and "extremely likely" as indicating the same level of certainty.

In summary, although linguistic measures of certainty exist, they have received relatively little validation and have not been tested for use with sentiment analysis specifically. It is possible they can suffice as measures of sentiment certainty, but this remains unclear. Moreover, even if validated, they would nonetheless suffer from several methodological limitations. The Certainty Lexicon (CL) addresses these limitations.

[2] DICTION's scoring system uses a word count approach and then standardizes this count. Thus, the underlying word count approach remains intact but is standardized. However, the way in which DICTION standardizes its word counts is not explained and therefore a step-by-step explanation of its scores is not possible (Hart and Carroll 2015). Nevertheless, the word count approach DICTION uses treats these sentences as equally certain.

Table 1. Comparison of the Certainty Lexicon (CL) with DICTION and LIWC, Example 1.

Text Example	Certainty (CL)	Certainty (DICTION)	Certainty (LIWC)	Tentativeness (LIWC)
I've sorta disliked my experience with that brand.	1.96	24.32	.00	12.50
I've often disliked my experience with that brand.	6.50	24.32	.00	12.50

Notes: Higher values indicate higher levels of the construct.

Table 2. Comparison of the Certainty Lexicon (CL) with DICTION and LIWC, Example 2.

Text Example	Certainty (CL)	Certainty (DICTION)	Certainty (LIWC)	Tentativeness (LIWC)
I'm not sure.	.97	74.94	33.33	.00
It seems like it could be fun.	4.34	21.88	.00	14.29
That is without doubt the worst.	8.58	29.11	.00	16.67

Notes: Higher values indicate higher levels of the construct.

The Certainty Lexicon

In this work, we construct the CL as a tool to quantify certainty in language. The CL is constructed to be a general measure of certainty in language—whether that certainty is about people's sentiment, thoughts, beliefs, or ideas. We validate the CL across a number of contexts, with a focus on assessing its ability to capture sentiment certainty.

At a conceptual level, the CL advances the sentiment analysis literature by recognizing and demonstrating the importance of moving beyond valence to understand sentiment and behavior. Beyond this conceptual advance, the CL addresses the methodological limitations of previous tools by using an approach called *imputation* (Rocklage, Rucker, and Nordgren 2018). Under an imputation approach, each word or phrase has its own normative certainty score that is empirically validated. Returning to the example of "often" versus "sorta," the current work finds that "often" has a normative certainty score of 6.50 out of 9.00 (0 = "very uncertain," and 9 = "very certain"). However, "sorta" has a much lower normative score of 1.96. Thus, "often" signals substantially greater certainty than "sorta." To quantify the amount of certainty in a piece of text, an imputation approach replaces each keyword with its normative score (i.e., the word is imputed with the normative score). Thus, whereas word count approaches constrain all words in a given dictionary to indicate identical certainty, the imputation approach used for the CL is sensitive to their differences (see Table 1).

The CL also overcomes the issue that short pieces of text, such as tweets from Twitter, contain a limited number of keywords for analysis. Imputation can provide diagnostic information even when only a single word related to certainty is present. Word count approaches would conclude that the two sentences in Table 1 are equivalent in certainty, whereas the CL can differentiate between them on the basis of one keyword.

Finally, although both LIWC and DICTION can only process single words in isolation, the CL takes phrases into consideration and therefore allows for a more accurate and nuanced measure of certainty. For example, LIWC and DICTION score the sentence "I'm not sure" high in certainty because they only process the word "sure." However, the CL would account for the phrase "not sure" and score this sentence correctly as indicating low certainty (see Table 2). Similarly, the sentence "that is without doubt the worst" is scored low in certainty for LIWC and DICTION because it contains the word "doubt." By contrast, the CL scores this high in certainty because it can consider the phrase "without doubt."

Overview

Our approach consists of three main stages: (1) we construct the CL as a measure of certainty in language across topics and contexts, (2) we validate its ability to assess sentiment certainty specifically, and (3) we demonstrate its utility in sentiment analysis. First, using a wide range of sources, we create a large corpus of words and phrases that individuals may use to convey their certainty. We obtain normative certainty scores from a large sample of external participants, and then empirically filter this extensive list to retain only words and phrases that provide consistent signals of certainty across diverse topics and contexts. This word list and its accompanying scores is the first of its kind and resulted in the CL. Second, we empirically validate the CL using both experiments and real-world data. Finally, we show that the CL offers insights beyond traditional sentiment analysis by analyzing tweets to predict future consumer behavior. Across these final stages, we also compare the CL with existing tools (LIWC and DICTION).

Constructing the Certainty Lexicon

Phase 1: Generating the Candidate Word List

We sought to create an extensive list of candidate words and phrases that people might use to communicate certainty across a wide range of topics (Table 3 summarizes the process).

Step 1: Automated extraction of words and phrases. In the first experiment, we elicited natural language from participants

Table 3. Summary of the Number of N-Grams Added or Removed at Each Step.

	Number of Words/Phrases	Details of Addition/Removal
Phase I: Generating Candidate Word List		
Initial word list (Steps 1–3)	1,425	Generated through two empirical studies and existing word lists
N-gram propagation (Step 4)	+35,193	Propagated synonymous words and phrases for each n-gram in the initial word list
Resulting word list	36,618	
Phase II: Initial Filtering		
Based on real-world frequency (Step 1)	−15,512	Removed n-grams that occurred with low frequency in news articles and Reddit data
Based on human assessment (Step 2)	−16,002	Removed n-grams judged as unknown or unlikely to be indicative of certainty
Resulting word list	5,104	
Phase III: Quantifying Certainty		
Based on human assessment	−1	Removed one n-gram that raters judged as unknown
Resulting word list	5,103	
Phase IV: Final Refinement		
Based on real-world prediction	−1,618	Removed n-grams where the normative score and regression coefficient were not directionally consistent
Final CL word list	3,485	

and used a data-driven approach to extract certainty-related words and phrases. To do so, we asked participants to describe their thoughts and feelings of a product/service, decision, or event that they were certain or uncertain about (see Web Appendix B). These three scenarios were selected for generalizability. A total of 612 participants from Amazon Mechanical Turk (MTurk) (61% female, 39% male; $M_{age} = 34.25$ years, $SD = 11.48$) were randomly assigned to conditions in a 2 (certain or uncertain) × 3 (scenario: product/service, decision, or event) between-participants design. Participants then reported their certainty about the scenario (1 = "very uncertain," and 7 = "very certain"; 1 = "very unsure," and 7 = "very sure"; $r = .95$, $p < .001$).

From participants' written text, we used differential language analysis (DLA) to identify the words and phrases—also known as *n-grams* in computational linguistics—that were most related to certainty and uncertainty (see Kern et al. 2016; Schwartz et al. 2013). DLA uses a data-driven approach to identify n-grams that are most strongly associated with a given construct (in this case, certainty). We describe the full DLA procedure in Web Appendix B. We extracted all n-grams from 1-grams to 4-grams (i.e., single words to four-word phrases).[3] There were 161,418 unique n-grams from the 612 participants. Given that our goal at this stage was to generate many candidate n-grams, we included an n-gram in the candidate list if it predicted participants' certainty at $p \leq .01$. This resulted in 122 words and phrases for the word list (e.g., "I was not," "unsure," "whether or not").

Step 2: N-gram listing. Next, we took a more direct approach to elicit n-grams related to certainty. Seventy participants in the United States (63% female, 37% male; $M_{age} = 36.06$ years, $SD = 13.41$), collected via MTurk, were asked to list words and phrases that conveyed either certainty or uncertainty. This step resulted in 397 unique n-grams. To extend the list of 1-grams further, we used synsets from WordNet (Miller 1995), which is a database that contains 155,327 total 1-grams and their synonyms. Using these synsets, 254 new 1-grams were added to the list.

Step 3: Supplemental sources. We also expanded the candidate word list using existing word lists that could plausibly relate to certainty. Three sources met this criterion.

First, we included the BioScope corpus, which is a collection of biomedical texts that have been annotated for how speculative different n-grams within scientific texts are (Vincze et al. 2008). Although the word list from BioScope is specific to biomedical texts, we reasoned that at least some of the identified words might generalize to other domains. For example, "likely" and "seems" were labeled as words indicating speculation and hedging. From this list, 100 new n-grams were added to our word list.

Second, we included word lists from DICTION (version 7.1; Hart and Carroll 2015), which contains only 1-grams. Three word lists from this source appeared potentially relevant for the certainty construct: tenacity, leveling, and ambivalence. These word lists are used in DICTION's measure of certainty. In total, this step added 281 new 1-grams to our word list.

Last, we included the "certainty" and "tentativeness" word lists from LIWC (Pennebaker et al. 2015), which also contain only 1-grams. For words that were stemmed in the original LIWC list (e.g., accura*), we completed the word in as many ways as possible using the "related forms" section of Dictionary.com. This led to the inclusion of 210 additional 1-grams.

[3] We also attempted to extract and analyze 5-grams. However, 5-grams that were significantly related to the certainty construct were exceedingly rare.

In total, Steps 1–3 generated a candidate word list of 1,425 n-grams.

Step 4: Seed n-gram propagation. Following prior work, we expanded this word list by using seed n-gram propagation, which identifies synonymous n-grams (Rocklage, Rucker, and Nordgren 2018). Specifically, we used two sources to generate synonymous n-grams. First, as in Step 2, we used synsets from WordNet (Miller 1995). We extracted all synsets for the 1,109 1-grams in the seed list.

Second, given that synsets include only 1-grams, we used an additional database to identify new n-grams beyond 1-grams. Specifically, we used the Paraphrase XL Database, a database of n-grams with 36.4 million words and phrases that contains synonymous words or phrases for n-grams (Pavlick et al. 2015). We allowed for plural nouns, singular nouns, and all tenses to be generated.

In total, 35,193 n-grams from WordNet and Paraphrase XL were added to the candidate list. Taken together, the n-gram generation steps (Phase I) resulted in a candidate word list with 36,618 unique words and phrases.

Phase II: Initial Filtering

We used a two-step procedure to refine the candidate word list from Phase I. First, we filtered the n-grams based on their frequency in real-world text. This enabled us to remove n-grams that were exceedingly rare and therefore unlikely to contribute to measuring certainty. Second, we had participants judge how well the remaining n-grams expressed certainty to further exclude n-grams that were unlikely to signal certainty. We describe each step in more detail.

Step 1: Filtering based on frequency of real-world use. We first aimed to filter out n-grams that occurred with low frequency in real text data (e.g., nonsense words, uncommon misspellings). To ensure that the final word list could apply to both more formal and informal language, we used two different sources of real-world language: news articles and Reddit posts.

To capture more formal language, we used news articles from the three top news outlets in the United States based on circulation (Cision Media Research 2019): the *New York Times*, the *Wall Street Journal*, and *USA Today*. We obtained all possible articles published by these outlets from 2010 to 2018 (Davies 2019). There was a total of 32,435 articles over this time. These articles covered a wide range of topics: art, politics, business, world news, sports, movies, technology, opinion, obituaries, and so on. The articles contained a total of 24.8 million words.

To capture informal language, we used posts on Reddit, one of the most visited websites on the internet (Alexa 2021), where people post comments and reactions to various topics. Specifically, we obtained every post on Reddit from every other month in 2018 (i.e., January, March, May, and so on), which was the most recent full year data were available (Baumgartner et al. 2020). We used six months of data to make data processing more manageable given the volume of data. These data contained 17.4 billion words in total from 613.4 million posts made by 11.6 million unique users. There were 187,081 unique "subreddits," which cover a wide array of topics such as investing, news, marketing, and sports.

From the news articles and Reddit posts, we extracted all possible n-grams (where n ranged from 1 to 4). Across the 10 years of news articles, 15.2 million unique n-grams were used on average per news outlet (SD = 3.06 million). These n-grams were used 19.7 million times in total. Across the six months of Reddit posts, 2.07 billion unique n-grams were used on average per month (SD = 61.3 million). These n-grams were used 56.2 billion times in total.

For each n-gram in the candidate word list, we counted how frequently that n-gram occurred in the news articles and Reddit posts. We then filtered out the n-grams that occurred with relatively low frequency. Specifically, n-grams that occurred fewer than 5,000 times across all sources were filtered out. Note that n-grams in our candidate word list were used 220,738 times on average across these sources (range = 0–803,155,564). Thus, this cutoff struck a balance between retaining the majority of the n-grams while filtering out those that were clearly rare and unlikely to be encountered in a vast majority of text (e.g., uncommon phrases, typos, misspellings such as "he's doubts" and "favorred"). This procedure removed 15,512 n-grams (42% of the candidate word list). We further removed 100 n-grams that fell into one of three categories: HTML formatting in the text data (e.g., br, p, h), numerals (e.g., 1, 2, 3), and single letters (e.g., b, c, e). This left us with a list of 21,106 n-grams.

Step 2: Filtering out words unrelated to certainty using human raters. Next, we refined the word list by having human raters judge the extent to which each of the 21,106 n-grams could be used to communicate certainty. The goal of this step was to filter out n-grams that were unlikely to communicate certainty. We aimed to have 10 participants judge each n-gram and determined based on prior research that each participant could feasibly rate 300 n-grams (Rocklage, Rucker, and Nordgren 2018). This indicated that approximately 704 participants were required (21,106 n-grams × 10 judgments per n-gram/300 judgments per participant). Seven hundred twenty-three participants completed this study on MTurk. We filtered out 41 participants whose first language was not English (i.e., they responded "no" to "Is English your first language?"), which left 682 participants in the data set (50% female, 50% male; M_{age} = 38.85 years, SD = 12.07).

Upon beginning the survey, participants were told that the researchers were interested in the language people use to express their certainty. Participants were then presented with the definition of certainty following prior research (Kagan 1972; Milliken 1987; Tormala and Rucker 2007) and examples of words or phrases that people may use to communicate their certainty (for details, see Web Appendix B). The instructions also clarified that some words may be unrelated to certainty (e.g., "refreshed," "I waited," "upgrades").

To ensure that participants understood the instructions, a practice session was administered prior to the judgment task (see Web Appendix B). Participants rated practice n-grams on the extent to which they could be used to indicate a sense of certainty or uncertainty (1 = "not at all," 3 = "somewhat," and 5 = "a great deal"). The instructions were made clear that participants' task was to judge the extent to which an n-gram was related to certainty (not the amount of certainty that an n-gram conveyed). Participants then rated approximately 300 randomly selected words from the list using the same scale from the practice session. Participants could also indicate that they did not know a word, given that their judgment would likely be inaccurate in those cases. Across all participants, a total of 205,556 judgments were made.

We retained n-grams that received an average score of 3 or higher (i.e., at least "somewhat" related to certainty or uncertainty). We then removed three n-grams that more than 50% of participants indicated they did not know ("nos," "tc," "noes"). The filtered list contained 5,104 n-grams (24% of the initial list).

Phase III: Quantifying the Certainty of Each N-Gram

Obtaining normative ratings. Next, we obtained normative ratings for each n-gram. These ratings form the basis of the Certainty Lexicon (CL) because they will be imputed each time the n-gram is used in a piece of text. Following prior work (Rocklage, Rucker, and Nordgren 2018), we aimed to have approximately 30 participants rate each n-gram. Given that there were 5,104 n-grams, this step required a sample size of approximately 510 participants (5,104 n-grams × 30 judgments per n-gram/300 judgments per participant). The final sample consisted of 515 participants from MTurk. As in the prior step, we filtered out 26 participants who indicated that their first language was not English. This left a final sample of 489 participants (50% female, 50% male; $M_{age} = 41.74$ years, $SD = 13.45$). These participants made a total of 146,783 judgments.

To elicit certainty ratings from participants, we provided them with the same definition of certainty as in the Phase II, Step 2 filtering study. Participants then completed a short practice session where they rated the implied certainty of each of these four n-grams on a scale from very uncertain to very certain (0 = "very uncertain," and 9 = "very certain"; see Web Appendix B). After completing the practice trials, each participant rated approximately 300 randomly selected n-grams using the same scale as in the practice trials. Based on these ratings, we removed one word because more than 50% of participants indicated they did not know it ("nt").

The normative certainty scores. To obtain the normative certainty score for each n-gram, we averaged participants' ratings for each n-gram. These scores captured the range of possible certainty, from high to low: "beyond any doubt" (M = 8.81), "it seems so" (M = 5.63), and "just don't know" (M = .63).[4]

Phase IV: Refining the Word List Using Real-World Data

Next, we empirically refined the word list using naturalistic text to retain just those words that provide a diagnostic signal of certainty in the real world. Indeed, one of the unique benefits of the CL is that each n-gram is empirically assessed for whether it predicts certainty in real-world text, whereas existing tools are often created by researchers manually composing a list of words they believe might indicate certainty (e.g., Hart and Carroll 2015; Pennebaker et al. 2015).

To assess the ability of each n-gram to predict certainty, we required real-world text that also included a quantitative measure of certainty. As an analogy, researchers often rely on online reviews (e.g., from Amazon, TripAdvisor, Yelp) as a source of real-world data to refine traditional sentiment analysis tools, which focus on valence (Liu 2012; Rocklage, Rucker, and Nordgren 2018). Online reviews are pivotal for constructing tools that measure valence because they contain naturalistic text along with a self-reported quantitative measure of consumers' attitude (i.e., a star rating). Researchers can therefore use these quantitative star ratings to examine whether, for example, the word "amazing" more often accompanies five-star or one-star reviews.

Given the widespread availability of such data, it is relatively straightforward to empirically assess whether a given n-gram provides a signal of positivity or negativity. However, it is more challenging to find a large collection of data that contains naturalistic text accompanied by a quantitative measure of individuals' certainty. To address this challenge, we turned to a novel source of data: online prediction markets.

Data. We obtained data from the website Good Judgment Open,[5] a prediction market open to anyone. The predictions cover topics across a wide variety of domains, including politics, finance, sports, health, technology, and entertainment. For example, individuals can submit a numeric probability for whether the annual sales of electric vehicles will reach a certain number by a given date in the future or whether a specific player will be signed to a future contract in the National Football League in the United States. Along with their numeric probability forecast, individuals can also write about why they issued the probabilities that they did. Thus, individuals' certainty is captured both via a numerical probability distribution and in linguistic form. As such, this prediction market

[4] Following prior research (Rocklage, Rucker, and Nordgren 2018), we also calculated the consistency of these certainty ratings for the final sample of 3,485 n-grams. Specifically, we randomly selected half of the participants and then calculated the average certainty for each of the n-grams from that half of the participants. We repeated this process an additional 99 times, each time using the full set of participants (i.e., sampling with replacement across the samples). Thus, there were 100 samples that contained the average certainty for each of the 3,485 n-grams from a randomly selected subset of participants. We correlated the 100 samples with each of the others (4,950 possible pairings). As evidence of the consistency of these ratings, there were strong correlations across the samples ($r_{avg} = .922$, 95% CI: [.921, .922]).

[5] https://www.gjopen.com/.

provides the necessary features—a quantitative measure of certainty and associated natural language—to empirically refine the word list to those n-grams that consistently signal higher versus lower certainty.

Each forecasting question prespecifies a set of potential outcomes. When making their predictions, individuals assign a probability (e.g., 50%) to each potential outcome. The questions on the website include binary yes/no questions ("Will Justin Trudeau cease to be prime minister of Canada after the next federal election?"; $n = 667$) and continuous questions that provide a range of options ("What will be the average number of Bitcoin transactions per day in the first week of June 2016?"; $n = 308$). For continuous questions, the website provides individuals with a prespecified range of ordinal outcomes (e.g., "Less than 125,000," "Between 125,000 and 250,000, inclusive," and "More than 250,000"; for additional details, see Web Appendix B). When submitting a forecast, individuals indicate the probability with which they believe each prespecified outcome will occur. The website requires that the assigned probabilities always sum to 100% across the prespecified outcomes. As we detail, these probability distributions can be transformed into a quantitative measure of individuals' certainty about the potential outcome.

When submitting a forecast, individuals can also write about why they issued the probabilities that they did. For example, when asked to predict whether more than a million refugees would arrive in Europe in 2016, one individual explained their assigned probability by writing, "The middle east is in crisis, and it is highly likely that citizens will continue to have to flee their country to find refuge."

We obtained all probability estimates that were accompanied by a text response from the launch of this prediction market in 2015 to 2020. There were 975 unique forecasting questions posed during this time, which garnered 169,954 sets of probability estimates from 20,793 unique individuals, thereby providing a large number of observations. As noted, the forecasting questions encompassed a wide range of topics, which helped ensure that the refined word list would provide a generalizable measure of certainty across topics.

Transforming probability forecasts into a quantitative measure of certainty. For each prediction, we quantified individuals' level of certainty about the outcome as the standard deviation of the probability distribution they specified. From Illowsky and Dean (2018):

$$\sqrt{\sum_{i=1}^{N} p_i(x_i - \mu)^2},$$

where p is the probability the individual assigned to a given outcome, x is the normalized range of possible answers (i.e., a number 0–1), and μ is the expected value (mean) calculated as

$$\sum_{i=1}^{N} p_i x_i.$$

A smaller standard deviation of individuals' probability distribution indicates greater certainty about which outcome will occur. Take, for example, an individual who forecasts a binary outcome. Individuals who assign a 90% chance to the outcome occurring (i.e., a probability distribution of 90%/10%) are significantly more certain about what will happen than those who assign a 50% chance to the outcome (i.e., a probability distribution of 50%/50%). The standard deviation captures this difference in certainty: the 90%/10% prediction has a standard deviation of .30, which is much smaller than the 50%/50% prediction's standard deviation of .50.

A similar intuition applies to the continuous forecasting questions, which specify more than two ordered outcomes (e.g., "Less than 125,000," "Between 125,000 and 250,000, inclusive," "More than 250,000"). Given that the prespecified outcomes are ordinal, the more certain individuals are of a particular outcome, the more concentrated their probability distribution should be (e.g., 90%/10%/0% reflects greater certainty than 40%/30%/30%), and thus the smaller the associated standard deviation would be. Put differently, if individuals believe that there could be relatively high probabilities of two extremely different outcomes (e.g., 40%/20%/40%), this indicates they have less certainty about what will happen than if they assign equally high probabilities to two adjacent outcomes (e.g., 20%/40%/40%). Thus, whereas the ordering of x makes no difference for calculating the standard deviation when there are only two outcomes, the ordering has meaning for the continuous questions, which have more than two ordered outcomes. We therefore coded each of the 308 continuous questions such that $x = 0$ represents the lowest option (e.g., "Less than 125,000") and $x = 1$ represents the highest option (e.g., "More than 250,000"). The middle option(s) (e.g., "Between 125,000 and 250,000, inclusive") received a normalized value between 0 and 1 (e.g., $x = .50$). Under this coding, the more concentrated the probability distribution (implying greater certainty), the smaller the resulting standard deviation.

Filtering methods and results. We then coded the text response associated with each forecast for whether it contained a given n-gram from our word list (coded as 1) or not (coded as 0). One dummy variable was generated for each of the 5,103 n-grams in the word list. Thus, each of the 169,954 observations was associated with 5,103 variables coded as 1 or 0.

Following prior work (Rocklage, Rucker, and Nordgren 2018), we then used each of these dummy variables in separate regressions to predict the standard deviation of the forecasts. An n-gram would have a positive coefficient if its presence systematically predicted a larger standard deviation (i.e., less certainty) and a negative coefficient if its presence systematically predicted a smaller standard deviation (i.e., greater certainty). We conducted separate regressions for each of the 5,103 n-grams for the binary questions and then, independently, for the continuous questions. Thus, two regression coefficients were obtained for each n-gram, resulting in a total of 10,206 regression coefficients ($5{,}103 \times 2$).

Using a similar approach as prior work (Rocklage, Rucker, and Nordgren 2018), we first filtered out n-grams that were not used in any of the text, because these n-grams are likely too rare to be useful (256 n-grams; 5% of the total). For the remaining n-grams, we retained an n-gram if at least one of its two regression coefficients was in the same direction as its normative score. Specifically, an n-gram was retained if its regression coefficient was negative for either the binary or continuous question type and its normative score was above the median normative certainty score (consistent signals of higher certainty), or if its regression coefficient was positive and its normative score was below the median normative certainty score (consistent signals of lower certainty).[6] This procedure filtered out 1,362 n-grams (27% of the total).

This process led to the retention of 3,485 n-grams that constitute the final list of the CL. For a summary of the number of words and phrases added or removed at each step of construction, see Table 3.

Validating the Certainty Lexicon

Having constructed the CL, the next stage was to validate it. Across a series of four studies, we used different empirical approaches to assess the validity of the CL. Study 1 examined how well the CL's measurement of certainty correlates with a measure of "ground truth" certainty. Study 2 provided an experimental assessment of the CL's ability to capture multiple levels of certainty, even when keeping the number of words constant. Then, we assessed the CL's specific ability to measure sentiment certainty using a large real-world data set of consumer reviews (Study 3). We then replicated these results experimentally (Study 4). These studies also assessed the performance of the CL compared with DICTION and LIWC.

Study 1: Comparing Measured Certainty with Ground Truth Certainty

Study 1 aimed to accomplish three objectives. First, we aimed to assess the CL's validity within natural text by examining its ability to capture differences in certainty. Second, we examined whether this validity held, even when accounting for the valence of consumers' sentiment. Finally, we sought to compare the results of the CL directly with those of DICTION and LIWC. In this experiment, we randomly assigned a large sample of participants to write a message to a friend about either something they were currently very certain of or something they were very uncertain of. The certainty of their text was measured using the CL, DICTION (version 7.1), and LIWC (Pennebaker et al. 2015). We then assessed how strongly each approach corresponded to ground truth as specified by the condition participants had been assigned to.

Procedure. Participants were randomly assigned to think of something they were very certain of or something they were very uncertain of. They were told that they would be writing a message to discuss a topic with a friend. To help ensure that participants wrote about a wide range of topics, we gave them four general domains to choose from: (1) a decision, (2) a future event, (3) a product or service, or (4) an issue. Choices were distributed among the topics as follows: decisions (34%), future events (34%), products/services (10%), and issues (22%). Those in the certain (uncertain) condition were then asked to write a message to a friend that explained their certainty (uncertainty) about their topic. Participants wrote an average of 104 words (SD = 51).

To measure CL certainty, we imputed the normative certainty of the CL n-grams participants used and then averaged these normative scores for each participant. For example, a participant whose text contained the CL phrases "I wasn't sure" (normative certainty: 1.76), "I think that" (4.61), and "not enough information" (1.24) would have a certainty score of 2.54 out of 9.00 ((1.76 + 4.61 + 1.24)/3). We used the standard "certainty" measure from DICTION and the "certainty" and "tentativeness" dictionaries from LIWC (Pennebaker et al. 2015).

To quantify traditional consumer sentiment (i.e., valence), we used a validated computational linguistic measure called the Evaluative Lexicon (Rocklage, Rucker, and Nordgren 2018). Specifically, we used the Evaluative Lexicon's measure of valence extremity (the deviation from the midpoint of the valence scale) to assess the degree of positivity or negativity. Using valence extremity is a conservative approach to examine the unique variance attributable to the CL given that prior research finds that more extreme valence can be associated with greater certainty, as discussed previously (e.g., Krosnick et al. 1993).

Participants. Participants were 978 individuals recruited via MTurk (57% female, 43% male; $M_{age} = 40.73$ years, SD = 13.85).

Results: DICTION and LIWC. We first examined the performance of DICTION and LIWC. To provide a standardized measure of how well these measures of certainty corresponded to ground truth, we calculated the point-biserial correlation between each of the measures and the condition participants had been assigned to (−1 = uncertainty, 1 = certainty). Point-biserial correlations provide a correlation coefficient between continuous (e.g., measured certainty) and dichotomous (e.g., condition) variables. The advantage of point-biserial correlations is that they can be interpreted in the same way as a traditional correlation coefficient (Howell 2016) and therefore enable us to compare the different methods on a familiar and easily interpreted metric.

[6] We relied on the median normative rating to separate n-grams into relatively high versus low certainty because it allowed us to retain the largest set of n-grams, therefore creating as comprehensive a list as possible. Alternative approaches to splitting the word list, such as using the mean certainty rating or the midpoint of the scale, resulted in less comprehensive word lists.

Table 4. Correlations Between Each Word List and Condition, Study 1.

	Certainty (CL)	Certainty (DICTION)	Certainty (LIWC)	Tentativeness (LIWC)	Certainty Index (LIWC)	Certainty (Condition)
Certainty (CL)	—					
Certainty (DICTION)	.26***	—				
Certainty (LIWC)	.20***	.38***	—			
Tentativeness (LIWC)	−.46***	−.25***	−.12***	—		
Certainty index (LIWC)	.46***	.39***	.62***	−.85***	—	
Certainty (condition)	.50***	.22***	.16***	−.35***	.36***	—

***$p < .001$.

Notes: Certainty (condition): −1 = uncertain condition, 1 = certain condition. Correlations between continuous measures and condition are point-biserial.

DICTION certainty exhibited a correlation strength of r = .22 ($p < .001$) with condition. LIWC certainty showed a correlation strength of r = .16 ($p < .001$). LIWC tentativeness showed a correlation of r = −.35 ($p < .001$). In line with prior conceptualizations of certainty as measured by LIWC (Pennebaker and Francis 1996), we also combined the two LIWC measures into a single index to form a continuous measure of high to low certainty (certainty minus tentativeness), which resulted in a correlation of r = .36 ($p < .001$). Thus, these correlations demonstrate the current state of measuring certainty in text and, perhaps ironically, provide some of the first direct evidence of the validity of these measures.

Results: Certainty Lexicon. We examined the results of the CL using the same correlation approach. CL certainty showed a correlation strength with condition of r = .50 ($p < .001$), which is approximately 2.3 times stronger than DICTION certainty, 3.1 times stronger than LIWC certainty, 1.4 times stronger than LIWC tentativeness, and 1.4 times stronger than the LIWC certainty index. Formally comparing the correlation coefficients (Fisher 1915), the CL correlation with ground truth was significantly stronger than the correlation for DICTION certainty ($Z = 7.42$, $p < .001$), LIWC certainty ($Z = 8.57$, $p < .001$), LIWC tentativeness ($Z = 4.06$, $p < .001$), and the LIWC certainty index ($Z = 3.81$, $p < .001$). In short, beyond providing the first formal test of validation of these certainty measures, the CL significantly outperformed all current text measures of certainty. For all correlations, see Table 4.[7]

We also assessed the CL's ability to measure certainty across topics. Attesting to the CL's generalizability, correlations between CL certainty and ground truth were of similar strength regardless of whether participants wrote about a decision (r = .49, $p < .001$), future event (r = .53, $p < .001$), product/service (r = .43, $p < .001$), or issue (r = .46, $p < .001$) (for the DICTION and LIWC analyses, see Web Appendix C). Furthermore, CL certainty continued to be strongly correlated with ground truth even when controlling for valence extremity in each topic: decisions (r = .47, $p < .001$), future events (r = .56, $p < .001$), products/services (r = .47, $p < .001$), and issues (r = .47, $p < .001$). This finding indicates that the CL explains additional variance beyond what traditional sentiment analysis captures with valence.

Discussion. Study 1 assessed the validity of the CL using naturalistic text in a controlled experimental setting. The CL was able to capture ground truth certainty and did so with greater accuracy than DICTION and LIWC, the predominant tools for measuring linguistic certainty. In an additional experiment, we also included a control condition. Results support the association between CL certainty and ground truth certainty (see Web Appendix E, Study W1).

Study 2: Sensitivity to Different Levels of Certainty

An important proposed strength of the CL is its sensitivity to different levels of certainty even when the number of words is constant. This ability is attributable to two features: (1) the use of imputation instead of word counts (i.e., each n-gram in the CL has its own certainty score) and (2) the inclusion of phrases, which help clarify the level of certainty (e.g., "very confident" vs. "confident" vs. "not confident"). DICTION and LIWC, however, use a word count approach and therefore treat all words as indicating the same level of certainty (for examples, see Table 2). Thus, unlike DICTION and LIWC, the CL should be able to distinguish between low, moderate, and high certainty even when the number of words does not vary. The purpose of this study was to provide a controlled assessment of this ability. To do so, we showed participants a subset of CL n-grams and asked them to select three n-grams that best communicated their high, moderate, or low level of certainty about a topic to others.

Procedure. Participants first read through a subset of 30 randomly selected n-grams from the CL that represented the full range of certainty (e.g., "absolutely sure," "it is possible that," "not clear"; the full list of 30 n-grams is provided in Web Appendix D). They were then asked to think of a product or service about which they were very certain (high certainty

[7] For robustness, we examined versions of the CL word list that used the mean or midpoint of the CL scale to filter the word list using the Good Judgment data. These versions provided very similar correlations with condition in the current study (r = .50 and r = .48, respectively).

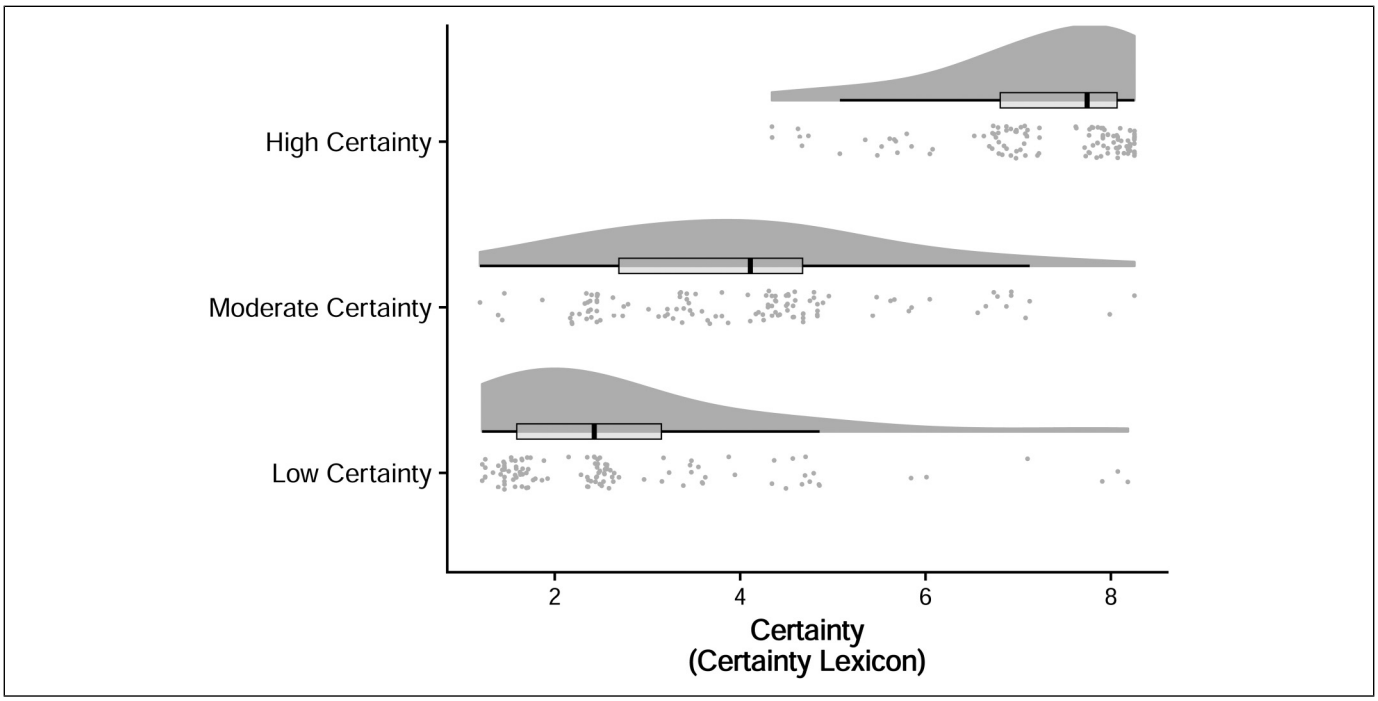

Figure 1. The Distribution, Boxplot, and the Associated Individual Data Points of CL Certainty Across Conditions, Study 2.

condition), somewhat certain (moderate certainty condition), or very uncertain (low certainty condition). Participants were told that they were going to write a message to a friend describing this product or service, and they were instructed to select three words or phrases from the list that they would use in their message to communicate how certain they were about that product or service. The dependent variable was the averaged certainty scores of participants' selected n-grams.

Participants. A sample of 356 Canadian undergraduate students completed this experiment in exchange for course credit (54% female, 46% male; $M_{age} = 21.56$ years, $SD = 3.85$).

Results. We examined the imputed certainty scores across the three conditions using a one-way analysis of variance. The effect of condition was significant ($F(2, 353) = 375.91$, $p < .001$, $\eta_p^2 = .68$). Imputed certainty was highest in the high certainty condition ($M = 7.27$, $SD = 1.02$), middling in the moderate certainty condition ($M = 3.97$, $SD = 1.46$), and lowest in the low certainty condition ($M = 2.66$, $SD = 1.47$). All conditions were significantly different from one another ($ps < .001$; Cohen's $d_{hi\ vs.\ mod} = 2.62$, $d_{mod\ vs.\ low} = .90$, $d_{hi\ vs.\ low} = 3.65$). Figure 1 illustrates the distribution of CL certainty scores across each condition.

Discussion. This simple validation experiment indicates that the CL is able to distinguish between different levels of certainty, even when the number of n-grams is held constant. Thus, whereas word count approaches such as DICTION and LIWC would treat individuals' responses as indicating the same level of certainty, the CL can differentiate between levels of certainty.

Study 3: Validating the Certainty Lexicon Using Real-World Data

We next aimed to validate the CL specifically as a measure of the certainty of consumer sentiment. Study 3 examined the CL in a conventional sentiment analysis context using naturalistic, real-world data: a large set of online product reviews. These consumer reviews enabled us to examine whether the CL provides outcomes that are consistent with prior research on attitude certainty and thereby assess the validity of the CL for measuring the certainty of consumer sentiment.

We examined two variables that existing research indicates should influence consumers' sentiment certainty. First, individuals feel less certain about their attitude when there is little existing attitude consensus among others (Cheatham and Tormala 2015; Luo, Raithel, and Wiles 2013; Maheswaran and Chaiken 1991; Tormala and Rucker 2018). In the context of online reviews, we predicted that consumers would express less certainty in their review if there was little existing consensus in other consumers' ratings of a product.

Second, research has demonstrated that people feel less certain when their attitude is at odds with the social consensus (Orive 1988; Petrocelli, Tormala, and Rucker 2007; Visser and Mirabile 2004). Indeed, prior work has found that individuals are more certain when their judgment is consistent with the group's average judgment (McGarty et al. 1993). Given that consumers often read prior reviews for a product before constructing their own reviews (Ludwig et al. 2013), we

investigated whether consumers expressed less certainty when their rating was more discrepant from the average rating expressed by others (i.e., the consensus).

Similar to Study 1, we further examined whether these two outcomes held even when controlling for the valence extremity of consumers' attitudes as measured by the Evaluative Lexicon. This approach enabled us to assess the robustness of the results and demonstrate the separable effects of sentiment certainty and valence.

These data also enabled us to conduct another comparison of the CL with LIWC. We focused on LIWC for three reasons. First, it is the most popular measure of linguistic certainty (Pennebaker et al. 2015). Second, after the CL, it showed the next highest correlation in the prior validation study. Third, the DICTION software was simply unable to process the text in the current study due to the large amount of data. Indeed, a limitation of DICTION is that it is less geared toward big data, thereby limiting its usefulness for sentiment analysis.

As noted, one of the strengths of the imputation approach is that it is less sensitive to the length of the text that is analyzed (Kteily et al. 2019; Rocklage and Rucker 2019). Given that online reviews have a large diversity of text lengths—some reviews are short, whereas others are quite long—we used this opportunity to examine the sensitivity of each approach to the length of the text. We hypothesized that LIWC would be more sensitive to length, whereas the CL would provide relatively consistent results regardless of text length.

Data. We obtained all online reviews from the Beer Advocate[8] website from its creation in 1996 to 2012, a span of 16 years (McAuley and Leskovec 2013). On this website, consumers review a beer they have tasted by issuing a quantitative star rating (one star through five stars) and writing text to explain their opinion. Each review includes the date consumers wrote their review, which enabled us to determine the order in which reviews were written about a given beer. There were 32,400 consumers who wrote a total of 1.43 million reviews that used at least one CL n-gram (out of 1.56 million reviews).

The Beer Advocate website displays both the average star rating and the variance of the existing star ratings; thus, consumers are presented with social consensus information when writing their review. To capture the degree of social consensus that each consumer would have seen, we calculated the standard deviation of ratings for all previous reviews for the beer up to when consumers wrote their review for that beer. Therefore, this was a time-varying index that changed for each additional review that was written about a beer. A higher standard deviation indicated less consensus about the quality of that beer ($M = .59$, $SD = .17$).

To quantify consumers' discrepancy from the social consensus, we calculated the absolute difference between each consumer's assigned rating for a given beer and the average of all previous reviews for the beer up to that date (abs(consumer's rating − running average rating)). Thus, this was also a time-varying index that captured consumers' discrepancy from the social consensus at the moment they wrote their review ($M = .47$, $SD = .42$).

Results: Certainty Lexicon. Given that some consumers wrote multiple reviews across beers, we used mixed-effects modeling to account for the variance attributable to each consumer (Nezlek 2011). We used the two variables of interest—the standard deviation of the prior ratings and the discrepancy of a user's rating from the existing average—to predict the user's certainty as measured by the CL. We also controlled for how many reviews had been written for the beer up to that point (log-transformed) given that early data points for a beer will naturally exhibit greater shifts in standard deviation and discrepancy from review to review. As more data points are accumulated for a beer, the standard deviation and discrepancy stabilize. Thus, accounting for this variable allows for greater generalization across the range of reviews and guards against the issue of having earlier observations overly influence the results. These three variables were used in the model to predict CL certainty.

As predicted, the less consensus there was about the beer up to the point at which a consumer wrote their review (i.e., the greater the standard deviation), the less certain they were ($\gamma = -.151$, $t(1,325,251.64) = 34.51$, $p < .001$). Similarly, the more discrepant consumers' ratings were from the social consensus to that point, the less certain they were ($\gamma = -.071$, $t(1,317,183.50) = 40.30$, $p < .001$). Both the consensus ($\gamma = -.137$, $t(1,325,220.68) = 31.11$, $p < .001$) and discrepancy ($\gamma = -.070$, $t(1,317,218.30) = 40.07$, $p < .001$) effects held when we also controlled for valence extremity. For additional analyses, see Web Appendix F. Note that standardizing coefficients for mixed-effects models can artificially affect within- and between-person variability (e.g., Blanton, Jaccard, and Burrows 2015). Due to this issue, the coefficients for these models are unstandardized, so the size of these coefficients cannot be directly compared within or between the different models.

Results: LIWC. We used the LIWC certainty index (certainty minus tentativeness) because it showed the strongest correspondence with ground truth certainty in the first validation study and also aligns with how certainty is conceptualized in prior research (i.e., a continuum of certainty to uncertainty [Pennebaker and Francis 1996]). For results of the LIWC certainty and tentativeness measures separately, which provide similar conclusions, see Web Appendix F.

We used the same mixed-effects model as in the previous section, now predicting the LIWC certainty index. Similar to the CL results, less consensus in prior ratings predicted the expected decrease in LIWC certainty ($\gamma = -.470$, $t(1,322,808.65) = 40.64$, $p < .001$). Discrepancy from social consensus, however, showed no significant relation with certainty as measured by LIWC ($\gamma = .007$, $t(1,324,867.83) = 1.41$, $p = .16$). The results for consensus ($\gamma = -.385$,

[8] https://www.beeradvocate.com/

$t(1,322,768.33) = 33.23$, $p < .001$) and discrepancy ($\gamma = .010$, $t(1,324,830.56) = 2.08$, $p = .038$) were similar when controlling for consumers' valence extremity.

Instability of LIWC results as a function of review length. As noted previously, one limitation of approaches such as LIWC is that they rely on word counts. These approaches can lead to a great deal of skewness in the data because a large number of texts, particularly those shorter in length, often receive scores of 0, indicating that LIWC words could not be detected in the text (Garten et al. 2018; Rocklage and Rucker 2019; Sterling, Jost, and Bonneau 2020). Corroborating this issue in the current data, shorter reviews (−1 SD) contained a median of zero LIWC certainty words and two LIWC tentativeness words, thereby providing a relatively poor signal of consumers' certainty. Longer reviews (+1 SD), on the other hand, contained a median of one LIWC certainty word and three LIWC tentativeness words.

To investigate whether this issue could explain the LIWC results, we conducted further analysis using those reviews that contained any signal of certainty (i.e., reviews with at least one LIWC certainty or tentativeness word) and moderated the effects of consensus and discrepancy by the length of each review (log-transformed). We hypothesized that LIWC would show similar results as the CL for longer reviews but discrepant results for shorter reviews.

There was both a consensus × review length ($\gamma = -.242$, $t(1,286,157.24) = 9.98$, $p < .001$) and a discrepancy × review length interaction ($\gamma = -.073$, $t(1,289,702.97) = 7.49$, $p < .001$). As hypothesized, for longer reviews (+1 SD), the LIWC results replicated those of the CL: less consensus ($\gamma = -.564$, $t(1,284,068.33) = 34.53$, $p < .001$) and greater discrepancy ($\gamma = -.018$, $t(1,288,738.70) = 2.79$, $p = .005$) both predicted less certainty. For shorter reviews (−1 SD), however, the results were inconsistent: less consensus predicted less certainty ($\gamma = -.338$, $t(1,288,553.09) = 20.59$, $p < .001$) but, contrary to expectations, greater discrepancy predicted more certainty ($\gamma = .050$, $t(1,287,094.63) = 7.50$, $p < .001$).

CL certainty, on the other hand, showed relatively little difference across reviews. Less consensus predicted significantly less certainty for both longer ($\gamma = -.153$, $t(1,323,685.48) = 24.92$, $p < .001$) and shorter ($\gamma = -.164$, $t(1,325,301.32) = 27.51$, $p < .001$) reviews. Similarly, discrepancy exhibited consistent effects for both longer ($\gamma = -.041$, $t(1,324,761.60) = 17.17$, $p < .001$) and shorter ($\gamma = -.093$, $t(1,314,395.07) = 37.76$, $p < .001$) reviews.

Discussion. These results validate the CL in a naturalistic setting. We found that the CL's measure of certainty was sensitive to variables that, based on prior research, should affect consumers' certainty. These results also held beyond valence extremity.

Moreover, whereas LIWC provided conflicting results for shorter versus longer pieces of text, the CL provided consistent results regardless of length. As noted previously, this difference is attributable to at least two key factors. First, the CL is fundamentally different in its measurement approach. Given that the CL uses imputation, whereas LIWC relies on word counts, the CL is more sensitive to differences in the certainty of individual words. For example, whereas LIWC treats the words "unknown" and "usually" as indicating the same level of certainty (both are from the tentativeness dictionary), the CL indicates that these words are associated with different levels of certainty (scores of 1.92 vs. 5.68, respectively). Second, the CL incorporates both words and phrases, and it contains a much more comprehensive dictionary that was generated and filtered using a data-driven approach designed to capture the construct of interest. Given these differences, the CL is able to provide a more precise measure of consumers' certainty and one that is also less volatile.

Study W2 provides another real-world validation of the CL in a different context (see Web Appendix G). Across approximately 18,000 consumers, we find that the more certain consumers are in their initial review of a restaurant on Yelp, the less their attitude changes when they revisit that restaurant in the future. These findings align with prior research that shows that greater certainty is associated with greater attitude strength (Petty and Krosnick 1995).

Study 4: Validation by Text Length

Study 4 had two major goals. First, following Study 3, we assessed the consistency of the CL experimentally by randomly assigning participants to write longer or shorter responses. We hypothesized that the CL would provide both stronger and more consistent correlations with ground truth certainty compared with LIWC, regardless of text length. LIWC, however, may show a particularly weaker correlation in shorter texts. Second, we aimed to extend the results of Study 1 by examining the CL's correlation with individuals' self-reported sentiment certainty.

Procedure. Participants were randomly assigned to think of an opinion that they were either very certain or very uncertain of. They were instructed to write a message to discuss their opinion with a friend. Participants could choose from three general domains: their opinion about a future event, a product or service, or an issue. Choices were fairly distributed among the topics: future events (37%), products/services (22%), and issues (41%). Participants were then randomly assigned to write either a longer message ("Please write your message so it is about one paragraph long") or a shorter message ("Please write just a few sentences"). The longer message condition produced messages that were approximately 40% longer on average: 67 words versus 49 words ($F(1, 280) = 21.42$, $p < .001$, $\eta_p^2 = .07$; for more, see Web Appendix H). Finally, participants reported their certainty about their opinion (1 = "very uncertain," 4 = "neither certain nor uncertain," and 7 = "very certain").

Participants. Participants were 284 individuals recruited via Prolific (49% female, 51% male; $M_{age} = 32.09$ years, SD = 11.65).

Results and discussion. CL certainty correlated with self-reported sentiment certainty of a similar strength as Study 1 ($r = .42$, $p < .001$). Moreover, it was consistent regardless of whether participants wrote about their opinion on a future event ($r = .40$, $p < .001$), product/service ($r = .44$, $p < .001$), or issue ($r = .40$, $p < .001$). Finally, the correlation was consistent across both long ($r = .42$, $p < .001$) and short ($r = .42$, $p < .001$) length conditions.

The correlation between the LIWC certainty index and self-reported sentiment certainty was weaker in strength ($r = .19$, $p = .002$). Moreover, this differed greatly depending on the topic participants wrote about: a future event ($r = .16$, $p = .10$), product/service ($r = .02$, $p = .88$), or issue ($r = .27$, $p = .004$). Finally, conceptually replicating the results from Study 3, the LIWC certainty index showed a significant association in the long length condition ($r = .24$, $p = .005$) but ultimately a nonsignificant association in the short length condition ($r = .15$, $p = .08$). For additional analyses and all correlations, see Web Appendix H.

Demonstration of Value: Social Media Sentiment Analysis

In this last study, we aimed to demonstrate both the importance of sentiment certainty for predicting consumers' future behavior and the real-world value of the CL. As discussed, sentiment analysis currently focuses on measuring the valence of individuals' attitudes (Lexico 2021). In this study, we show that measuring sentiment certainty improves the prediction of behavior.

To that end, we focused on a context of considerable importance to marketers: Super Bowl commercials. When marketers invest in a Super Bowl commercial, a primary aim is to drive greater awareness and engagement beyond the 30-second view time by, for example, attracting consumers to their brand on social media. In this study, we quantified both the valence and certainty of consumer sentiment toward Super Bowl commercials, as expressed in their real-time Twitter posts while they watched the Super Bowl. We then used these measures to predict one behavioral measure of commercial success: the number of new followers each brand gained on Facebook in the two weeks after the Super Bowl.

These data provided an interesting context to examine the CL for two reasons. First, prior research shows that people are relatively positive toward Super Bowl commercials (*USA Today* Ad Meter 2017), thereby creating a restricted range of positivity. This suggests that the conventional sentiment analysis approach may not be adequate. Second, Twitter data enable another examination of the CL's utility with short pieces of text. At the time of these Super Bowls (2016 and 2017), Twitter allowed a maximum of 140 characters per tweet and, indeed, the average tweet length in the current data is only 12 words. As noted previously, a unique benefit of the CL is that it can provide a more nuanced measure of consumers' certainty even with a single word. We also compare the CL results with LIWC.

Data

We obtained all real-time tweets about the commercials that ran during the 2016 and 2017 Super Bowls (Rocklage, Rucker, and Nordgren 2021b). To ensure that tweets were about a given commercial, not just about the company in general, we obtained tweets that (1) were posted during the time the commercials were running, (2) mentioned the name of the company or an affiliated keyword from the commercial itself (e.g., "MadeByGoogle"), and (3) specifically referenced either the Super Bowl or a commercial (for terms, see Web Appendix I). There were 94 commercials from 84 brands and 130,000 tweets about these commercials. To quantify consumer sentiment, we measured the valence of the tweets for each commercial using the Evaluative Lexicon.[9] To quantify sentiment certainty, we used the CL and LIWC. Thus, we measured the average valence and certainty consumers expressed toward each commercial.

As a metric of success, we recorded the daily number of new followers each brand accrued on Facebook in the two weeks after each Super Bowl (for each brand's Facebook page, see Web Appendix I). We also indexed the average number of daily followers brands accrued prior to the Super Bowl to assess the increase in the average number of followers for each brand after the Super Bowl. Given that Facebook did not provide an accessible historical record for brands' pages, we collected the number of followers for each brand as soon as it announced it would air a commercial during the Super Bowl. We collected an average of 21.85 days of daily new followers for each brand prior to the 2016 Super Bowl ($SD = 7.83$) and 16.05 days for the 2017 Super Bowl ($SD = 10.73$). We then extracted the daily number of new followers for each brand for the two weeks after each Super Bowl. The average number of daily new followers (log-transformed) in the two weeks after the Super Bowl was the dependent variable. Given that each brand had only a single Facebook page, valence and certainty were averaged across the commercials if a brand showed more than one commercial during a given Super Bowl.

Results

First, in line with prior results (*USA Today* Ad Meter 2017), we found that consumers were overwhelmingly positive toward the Super Bowl commercials ($M = 6.13$ out of 9.00). Only four commercials elicited valence that was negative (i.e., below the midpoint of the Evaluative Lexicon valence scale). Indeed, as prior research indicates, a "positivity problem" often exists such that most expressed consumer sentiment is positive (Rocklage, Rucker, and Nordgren 2021b). The current field study therefore reflects a common challenge marketers face when attempting to predict future behavior based on a restricted range of consumer sentiment.

[9] Before analyzing the data, we removed the word "super" from this word list given that there were many references to the "Super Bowl," which is a nonevaluative phrase in this data set.

Table 5. Regression Models Predicting Daily New Facebook Followers.

	CL Certainty (1)	Controls (2)	LIWC (3)
Primary Predictors			
Certainty (CL)	.024* (.01)	.022* (.01)	
Sentiment (EL valence)	−.007 (.01)	−.018 (.01)	−.003 (.01)
Sentiment (# pos vs. neg tweets)	−.011 (.01)	.006 (.01)	−.005 (.01)
Certainty index (LIWC)			.016 (.01)
Additional Control Variables			
Daily followers (pre–Super Bowl)	.141*** (.01)	.143*** (.01)	.143*** (.01)
Number of commercials		−.003 (.01)	
Super Bowl		.011 (.01)	
Quarter shown		−.007 (.01)	
Emotionality (EL)		.027* (.01)	

*$p \leq .05$.
***$p \leq .001$.
Notes: Super Bowl: −1 = Super Bowl 2016, 1 = Super Bowl 2017; all other predictor variables are standardized. Standard errors in parentheses. EL = Evaluative Lexicon.

Using regression, we predicted the number of new Facebook followers each brand gained based on the valence and certainty of the tweets about each commercial. We also included the number of positive versus negative tweets about each commercial given that prior work shows the relevance of this variable in sentiment analysis (Asur and Huberman 2010; O'Connor et al. 2010). Finally, we controlled for the average number of daily new followers each brand had in the days prior to each Super Bowl (log-transformed).

As we might expect, the more daily followers brands gained prior to the Super Bowl, the more they accrued in the two weeks after ($B = .141$, $t(79) = 14.45$, $p < .001$). Valence ($B = -.007$, $t(79) = .59$, $p = .56$), and the number of positive versus negative tweets ($B = -.007$, $t(79) = .99$, $p = .32$) were nonsignificant predictors.[10] However, the certainty consumers expressed in these very same tweets predicted future Facebook followers: greater certainty predicted that the company would accrue more followers two weeks later ($B = .024$, $t(79) = 2.30$, $p = .02$).[11]

To examine whether certainty was a unique predictor, we also assessed its robustness with additional controls. We controlled for the number of commercials each company showed during each Super Bowl, which Super Bowl the commercial was shown during (2016 vs. 2017), and the quarter of the game in which the commercial was shown. We also controlled for the emotionality of consumers' tweets as measured by the EL, which has been shown to predict consumer behavior in prior research (Rocklage, Rucker, and Nordgren 2021b). Certainty remained a significant predictor beyond these controls (see Table 5).

Finally, we assessed the ability of LIWC to predict future Facebook followers. The LIWC certainty index was not a significant predictor (see Table 5).[12]

Discussion

These results show that consumer certainty acts as a unique signal of future behavior. Moreover, they demonstrate that certainty can increase predictive ability in real marketing situations where traditional sentiment analysis cannot—in the present context, valence did not predict future behavior, likely because brands make largely favorable Super Bowl commercials, which restricts the range of valence. These results also demonstrate that the CL can be used by marketers in situations where pieces of text are short, which is an exceedingly common need for practitioners. LIWC, on the other hand, was not a significant predictor.

General Discussion

Analyzing the language and judgments of over 11.6 million people across contexts ranging from news articles, Reddit posts, prediction markets, online reviews, Twitter posts, and lab experiments, the current work introduces the first validated measure of certainty for use with sentiment analysis: the Certainty Lexicon (CL).

This work makes three main contributions. First, it makes a conceptual advancement by pushing sentiment analysis beyond its focus on valence. Research on attitudes has observed that although valence alone can be an unreliable predictor of future attitudes and behavior (Petty and Krosnick 1995; Wicker 1969), incorporating certainty can help predict a range of important

[10] These null results are not specific to the EL's measure of valence. Indeed, greater positive versus negative valence as measured by LIWC was also a nonsignificant predictor ($B = -.003$, $t(81) = .33$, $p = .74$).

[11] Although one might predict a valence × sentiment interaction, this interaction was nonsignificant given the small number of negative commercials ($B = -.006$, $t(79) = .56$, $p = .58$).

[12] Similarly, neither LIWC certainty ($B = .008$, $t(78) = .74$, $p = .46$) nor LIWC tentativeness ($B = -.016$, $t(78) = 1.48$, $p = .14$) were significant predictors when entered into the model instead of the LIWC certainty index.

outcomes (Tormala and Rucker 2018). We introduce these conceptual advances to the measurement of consumer sentiment and offer the first empirically derived and validated linguistic measure of certainty for use with sentiment analysis.

Second, for a field fixated on sentiment valence, this research shows the consequential value of measuring sentiment certainty. We found that certainty has predictive ability beyond valence, and it predicted behavior that valence did not. Whereas the valence of tweets on Twitter did not predict a Super Bowl commercial's future success, the certainty expressed in these same tweets did. Moreover, as reported in Web Appendix G, across approximately 18,000 consumers, we found that the more certainty consumers expressed in their initial review of a restaurant on Yelp, the less their attitude changed when they revisited the restaurant. Indeed, the predictive utility of certainty was about as strong as the extremity of the attitude itself.

Finally, this work offers methodological advances. We use modern developments in computational linguistics coupled with big data to construct the CL, thereby providing a road map for the construction and validation of future linguistic measures. This work also compared the validity of the CL with existing measures of certainty from LIWC (Pennebaker et al. 2015) and DICTION (Hart and Carroll 2015). We show that the CL provides greater measurement accuracy and more reliable insights into certainty and its effects.

Using the Certainty Lexicon

We constructed and validated the CL to measure certainty across contexts. In this work, we pay particular attention to validating the CL as a measure of sentiment certainty (i.e., certainty associated with attitudes). Nevertheless, the CL is a general measure of certainty in language, and our validation tests demonstrate its generalizability. For example, the CL can be used to capture differences in individuals' certainty about future events and decisions (see Study 1) that are not inherently sentiment focused. Thus, as a natural language tool devoted to certainty, researchers can apply the CL to different contexts and not only for sentiment analysis.

When performing sentiment analysis, researchers can now assess both sentiment valence and certainty. In some cases, valence and certainty may interact to predict behavior (e.g., Luttrell, Petty, and Briñol 2016). For example, consumers who are both positive and certain may be more likely to repurchase a product, whereas those who are positive and uncertain may be less likely to repurchase. When valence has a restricted range, which is often the case with consumer sentiment expressed online (Rocklage, Rucker, and Nordgren 2021b), certainty may be a better predictor than valence. Our Super Bowl study provides an example of this. Thus, best practice would be to assess both sentiment valence and certainty.

Additional Implications and Future Directions

The present work has several implications for researchers interested in understanding attitude certainty. Nearly all the attitude certainty findings in consumer behavior and psychology rely on self-report measures taken in the lab. The reliance on self-report can be a significant limitation given the marketing discipline's focus on practical, real-world contexts and applications. The CL can be used to test hypotheses using naturalistic data.

Similarly, lab findings involving certainty can now be applied in real-world contexts. For example, prior research suggests that people who are uncertain of an attitude may be more persuadable (Tormala and Petty 2002). Thus, marketers might track changes in sentiment certainty as an indication that it is time to advertise to bolster the attitudes of their customers. Whereas valence might lead to the conclusion that all is well, sentiment certainty may reveal consumers are in danger of changing their attitudes.

Most research in marketing concentrates on individual observations from a given consumer, such as a single online review, one tweet, or one Facebook post. It is rare to see these individual observations connected across time. Yet, many consumers write multiple reviews or post numerous tweets and Facebook entries across time. The CL enables researchers to quantify large amounts of data efficiently and accurately and thus to trace consumers' certainty over time. This can enable researchers to, for example, understand how consumer certainty changes as consumers gain expertise in a domain and the implication of these changes for understanding consumer behavior (e.g., Rocklage, Rucker, and Nordgren 2021a). Do consumers become more certain in their attitude as their expertise develops, or do they become less certain because they have a more nuanced understanding of the domain?

On a more macro scale, the CL can enable researchers and practitioners to track fluctuations in certainty across time via sources such as news articles or social media. For example, Apple is often a master at creating uncertainty and anticipation around their product launches, which is likely to be reflected in consumers' language prior to these launches. Sometimes, however, the launches themselves fall short of expectations and disappoint consumers (Sherman 2011). Companies may consider tracking this uncertainty to strike a balance between keeping consumers' attention while not generating anticipation beyond what the product can support.

This research also opens numerous possibilities for examining and understanding how expressed certainty might affect other consumers. How does certainty expressed by one consumer affect others when it comes to word of mouth? As suggested in prior research, linguistic certainty is likely to have implications for the impact of advertising (Briñol, Petty, and Tormala 2004), the spread of information (Dubois, Rucker, and Tormala 2011), and for consumers who advocate for a product or cause (Barden and Petty 2008).

Of note, because the CL was constructed as a general measure of certainty in language, it has the potential for broader application to other contexts in which gleaning certainty from language can offer insights. Beyond certainty related to attitudes, the CL may be used to measure certainty that people express about beliefs, future events, or topics in a variety of contexts.

Practitioners can also make use of the CL list itself. Similar to prior work showing the importance of word selection (e.g., Sela, Wheeler, and Sarial-Abi 2012), the CL can be used by practitioners to carefully select the degree of certainty they wish to convey. Though expressing high certainty in copy is often effective, expressing some degree of uncertainty can nudge consumers to think more about the product (Reich and Tormala 2013).

Conclusion

The current work highlights the importance of moving beyond traditional sentiment analysis toward the measurement of sentiment certainty. We provide researchers and practitioners the opportunity to explore the impact of certainty both in the lab and "in the wild" via the Certainty Lexicon—available for download, free of cost, at www.CertaintyLexicon.com.

Associate Editor

Aner Sela

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Alexa (2021), "Alexa Top Sites," (accessed January 27, 2021), https://www.alexa.com/topsites.

Asur, Sitaram and Bernardo A. Huberman (2010), "Predicting the Future with Social Media," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, 492–99.

Babić Rosario, Ana, Francesca Sotgiu, Kristine De Valck, and Tammo H.A. Bijmolt (2016), "The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors," *Journal of Marketing Research*, 53 (3), 297–318.

Barden, Jamie and Richard E. Petty (2008), "The Mere Perception of Elaboration Creates Attitude Certainty: Exploring the Thoughtfulness Heuristic," *Journal of Personality and Social Psychology*, 95 (3), 489–509.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn (2020), "The Pushshift Reddit Dataset," in *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. Association for the Advancement of Artificial Intelligence, 830–39.

Berger, Jonah (2014), "Word of Mouth and Interpersonal Communication: A Review and Directions for Future Research," *Journal of Consumer Psychology*, 24 (4), 586–607.

Blanton, Hart, James Jaccard, and Christopher N. Burrows (2015), "Implications of the Implicit Association Test D-Transformation for Psychological Assessment," *Assessment*, 22 (4), 429–40.

Briñol, Pablo, Richard E. Petty, and Zakary L. Tormala (2004), "Self-Validation of Cognitive Responses to Advertisements," *Journal of Consumer Research*, 30 (4), 559–73.

Burn-Murdoch, John (2013), "Social Media Analytics: Are We Nearly There Yet?" *The Guardian* (June 10), http://www.theguardian.com/news/datablog/2013/jun/10/social-media-analytics-sentiment-analysis.

Cheatham, Lauren and Zakary L. Tormala (2015), "Attitude Certainty and Attitudinal Advocacy: The Unique Roles of Clarity and Correctness," *Personality and Social Psychology Bulletin*, 41 (11), 1537–50.

Chen, Zoey and May Yuan (2020), "Psychology of Word of Mouth Marketing," *Current Opinion in Psychology*, 31 (February), 7–10.

Cision Media Research (2019), "Top 10 U.S. Daily Newspapers," (accessed January 27, 2021), https://www.cision.com/2019/01/top-ten-us-daily-newspapers/.

Clarkson, Joshua J., Zakary L. Tormala, and Christopher Leone (2011), "A Self-Validation Perspective on the Mere Thought Effect," *Journal of Experimental Social Psychology*, 47 (2), 449–54.

Dancshazy, Csaba (2021), "Microsoft Gains More Value from Social and Survey Data with Lexalytics," (accessed March 1, 2021), https://www.lexalytics.com/resources/Microsoft-Case-Study.pdf.

Davies, Mark (2019), "The Best of Both Worlds: Multi-Billion Word 'Dynamic' Corpora," in *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache, 23–28.

De Langhe, Bart, Philip M. Fernbach, and Donald R. Lichtenstein (2016), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*, 42 (6), 817–33.

Dubois, David, Derek D. Rucker, and Zakary L. Tormala (2011), "From Rumors to Facts, and Facts to Rumors: The Role of Certainty Decay in Consumer Communications," *Journal of Marketing Research*, 48 (6), 1020–32.

Fisher, Ronald A. (1915), "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," *Biometrika*, 10 (4), 507–21.

Franc, Renata (1999), "Attitude Strength and the Attitude–Behavior Domain: Magnitude and Independence of Moderating Effects of Different Strength Indices," *Journal of Social Behavior and Personality*, 14 (2), 177–95.

Garten, Justin, Joe Hoover, Kate M. Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani (2018), "Dictionaries and Distributions: Combining Expert Knowledge and Large Scale Textual Data Content Analysis," *Behavior Research Methods*, 50 (1), 344–61.

Google Books Ngrams (2021), "Sentiment Analysis," (accessed March 1, 2021), https://books.google.com/ngrams/.

Google Trends (2021), "Sentiment Analysis," (accessed March 1, 2021), https://trends.google.com/.

Hart, Roderick P. (1976), "Absolutism and Situation: Prolegomena to a Rhetorical Biography of Richard M. Nixon," *Communication Monographs*, 43 (3), 204–28.

Hart, Roderick P. (1984), "The Language of the Modern Presidency," *Presidential Studies Quarterly*, 14 (2), 249–64.

Hart, Roderick P. and Craig E. Carroll (2015), *DICTION 7.1 Help Manual*. Digitext, Inc.

Holbrook, Morris B. and Michela Addis (2007), "Taste Versus the Market: An Extension of Research on the Consumption of Popular Culture," *Journal of Consumer Research*, 34 (3), 415–24.

Howell, David C. (2016), *Fundamental Statistics for the Behavioral Sciences, 9th ed*. Cengage Learning.

Humphreys, Ashlee and Rebecca Jen-Hui Wang (2018), "Automated Text Analysis for Consumer Research," *Journal of Consumer Research*, 44 (6), 1274–306.

Illowsky, Barbara and Susan Dean (2018), *Introductory Statistics*. Open Access Textbooks.

Kagan, Jerome (1972), "Motives and Development," *Journal of Personality and Social Psychology*, 22 (1), 51–66.

Kern, Margaret L., Gregory Park, Johannes C. Eichstaedt, H. Andrew Schwartz, Maarten Sap, and Laura K. Smith, and Lyle H. Ungar (2016), "Gaining Insights from Social Media Language: Methodologies and Challenges," *Psychological Methods*, 21 (4), 507–25.

Kessler, Sarah (2014), "The Problem with Sentiment Analysis," *Fast Company* (November 5), https://www.fastcompany.com/3037915/the-problem-with-sentiment-analysis.

Krosnick, Jon A., David S. Boninger, Yao C. Chuang, Matthew K. Berent, and Catherine G. Carnot (1993), "Attitude Strength: One Construct or Many Related Constructs?" *Journal of Personality and Social Psychology*, 65 (6), 1132–51.

Kteily, Nour S., Matthew D. Rocklage, Kaylene McClanahan, and Arnold K. Ho (2019), "Political Ideology Shapes the Amplification of the Accomplishments of Disadvantaged vs. Advantaged Group Members," *Proceedings of the National Academy of Sciences*, 116 (5), 1559–68.

Lexico (2021), "Sentiment Analysis," (accessed March 1, 2021), https://www.lexico.com/en/definition/sentiment_analysis.

Litt, Ab and Zakary L. Tormala (2010), "Fragile Enhancement of Attitudes and Intentions Following Difficult Decisions," *Journal of Consumer Research*, 37 (4), 584–98.

Liu, Bing (2012), *Sentiment Analysis and Opinion Mining*. Springer.

Ludwig, Stephen, Ko de Ruyter, Mike Friedman, Elisabeth C. Brüggen, Martin Wetzels, and Gerard Pfann (2013), "More Than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates," *Journal of Marketing*, 77 (1), 87–103.

Luo, Xuming, Sascha Raithel, and Michael A. Wiles (2013), "The Impact of Brand Rating Dispersion on Firm Value," *Journal of Marketing Research*, 50 (3), 399–415.

Luttrell, Andrew, Richard E. Petty, and Pablo Briñol (2016), "Ambivalence and Certainty Can Interact to Predict Attitude Stability over Time," *Journal of Experimental Social Psychology*, 63 (March), 56–68.

Maheswaran, Durairaj and Shelly Chaiken (1991), "Promoting Systematic Processing in Low-Motivation Settings: Effect of Incongruent Information on Processing and Judgment," *Journal of Personality and Social Psychology*, 61 (1), 13–25.

McAuley, Julian John and Jure Leskovec (2013), "From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise Through Online Reviews," in *WWW '13: Proceedings of the 22nd International Conference on World Wide Web*. Association for Computing Machinery, 897–908.

McGarty, Craig, John C. Turner, Penelope J. Oakes, and S. Alexander Haslam (1993), "The Creation of Uncertainty in the Influence Process: The Roles of Stimulus Information and Disagreement with Similar Others," *European Journal of Social Psychology*, 23 (1), 17–38.

Mighty Guides (2016), "Text Analytics: Risk Mitigation with Text Analysis," (March 4), https://www.slideshare.net/DavidRogelberg/text-analytics-mini-book-risk-mitigation-with-text-analytics.

Miller, George (1995), "WordNet: A Lexical Database for English," *Communications of the ACM*, 38 (11), 39–41.

Milliken, Frances J. (1987), "Three Types of Perceived Uncertainty About the Environment: State, Effect, and Response Uncertainty," *Academy of Management Review*, 12 (1), 133–43.

Moore, Sarah G. and Katherine C. Lafreniere (2020), "How Online Word-of-Mouth Impacts Receivers," *Consumer Psychology Review*, 3 (1), 34–59.

Nezlek, John B. (2011), *Multilevel Modeling for Social and Personality Psychology*. SAGE Publications.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith (2010), "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, 122–29.

Orive, Ruben (1988), "Group Consensus, Action Immediacy, and Opinion Confidence," *Personality and Social Psychology Bulletin*, 14 (3), 573–77.

Pavlick, Ellie, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch (2015), "PPDB 2.0: Better Paraphrase Ranking, Fine-Grained Entailment Relations, Word Embeddings, and Style Classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, 425–30.

Pennebaker, James W., Ryan L. Boyd, Kayla Jordan, and Kate Blackburn (2015), "The Development and Psychometric Properties of LIWC2015," University of Texas at Austin (September 15), https://repositories.lib.utexas.edu/handle/2152/31333.

Pennebaker, James W. and Martha E. Francis (1996), "Cognitive, Emotional, and Language Processes in Disclosure," *Cognition and Emotion*, 10 (6), 601–26.

Petrie, Keith J., Roger J. Booth, and James W. Pennebaker (1998), "The Immunological Effects of Thought Suppression," *Journal of Personality and Social Psychology*, 75 (5), 1264–72.

Petrocelli, John V., Zakary L. Tormala, and Derek D. Rucker (2007), "Unpacking Attitude Certainty: Attitude Clarity and Attitude Correctness," *Journal of Personality and Social Psychology*, 92 (1), 30–41.

Petty, Richard E. and Jon A. Krosnick, eds. (1995), *Attitude Strength: Antecedents and Consequences.* Psychology Press.

Qin, Qi, Wenpeng Hu, and Bing Liu (2020), "Using the Past Knowledge to Improve Sentiment Classification," in *Findings of*

the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, 1124–33.

Reich, Taly and Zakary L. Tormala (2013), "When Contradictions Foster Persuasion: An Attributional Perspective," *Journal of Experimental Social Psychology*, 49 (3), 426–39.

Rocklage, Matthew D. and Andrew Luttrell (2021), "Attitudes Based on Feelings: Fixed or Fleeting?" *Psychological Science*, 32 (3), 364–80.

Rocklage, Matthew D. and Derek D. Rucker (2019), "Text Analysis in Consumer Research: An Overview and Tutorial," in *Handbook of Research Methods in Consumer Psychology*, Frank R. Kardes, Paul M. Herr, and Norbert Schwarz, eds. Routledge, 385–402.

Rocklage, Matthew D., Derek D. Rucker, and Loran F. Nordgren (2018), "The Evaluative Lexicon 2.0: The Measurement of Emotionality, Extremity, and Valence in Language," *Behavior Research Methods*, 50 (4), 1327–44.

Rocklage, Matthew D., Derek D. Rucker, and Loran F. Nordgren (2021a), "Emotionally Numb: Expertise Dulls Consumer Experience," *Journal of Consumer Research*, 48 (3), 355–73.

Rocklage, Matthew D., Derek D. Rucker, and Loran F. Nordgren (2021b), "Mass-Scale Emotionality Reveals Human Behaviour and Marketplace Success," *Nature Human Behaviour*, 5 (10), 1323–29.

Rucker, Derek D. and Richard E. Petty (2004), "When Resistance Is Futile: Consequences of Failed Counterarguing for Attitude Certainty," *Journal of Personality and Social Psychology*, 86 (2), 219–35.

Rucker, Derek D., Richard E. Petty, and Pablo Briñol (2008), "What's in a Frame Anyway?: A Meta-Cognitive Analysis of the Impact of One Versus Two Sided Message Framing on Attitude Certainty," *Journal of Consumer Psychology*, 18 (2), 137–49.

Rucker, Derek D., Zakary L. Tormala, Richard E. Petty, and Pablo Briñol (2014), "Consumer Conviction and Commitment: An Appraisal-Based Framework for Attitude Certainty," *Journal of Consumer Psychology*, 24 (1), 119–36.

Schaefer, Mark W. (2015), "Get More Value from 'Gray Social'," *Harvard Business Review* (April 29), https://hbr.org/2015/04/get-more-value-from-gray-social.

Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E.P. Seligman, and Lyle H. Ungar (2013), "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," *PLoS One*, 8 (9), e73791.

Sela, Aner, S. Christian Wheeler, and Gülen Sarial-Abi (2012), "We Are Not the Same as You and I: Causal Effects of Minor Language Variations on Consumers' Attitudes Toward Brands," *Journal of Consumer Research*, 39 (3), 644–61.

Seyeditabari, Armin, Narges Tabari, and Wlodek Zadrozny (2018), "Emotion Detection in Text: A Review," arXiv, https://doi.org/10.48550/arXiv.1806.00674.

Sherman, Erik (2011), "No iPhone 5: Apple Totally Blows Its Moment," *CBS News* (October 5), https://www.cbsnews.com/news/no-iphone-5-apple-totally-blows-its-moment/.

Sterling, Joanna, John T. Jost, and Richard Bonneau (2020), "Political Psycholinguistics: A Comprehensive Analysis of the Language Habits of Liberal and Conservative Social Media Users," *Journal of Personality and Social Psychology*, 118 (4), 805–34.

Stone, Philip J., Robert F. Bales, J. Zvi Namenwirth, and Daniel M. Ogilvie (1962), "The General Inquirer: A Computer System for Content Analysis and Retrieval Based on the Sentence as a Unit of Information," *Behavioral Science*, 7 (4), 484–98.

Tormala, Zakary L. (2016), "The Role of Certainty (and Uncertainty) in Attitudes and Persuasion," *Current Opinion in Psychology*, 10 (August), 6–11.

Tormala, Zakary L. and Victoria L. DeSensi (2009), "The Effects of Minority/Majority Source Status on Attitude Certainty: A Matching Perspective," *Personality and Social Psychology Bulletin*, 35 (1), 114–25.

Tormala, Zakary L. and Richard E. Petty (2002), "What Doesn't Kill Me Makes Me Stronger: The Effects of Resisting Persuasion on Attitude Certainty," *Journal of Personality and Social Psychology*, 83 (6), 1298–1313.

Tormala, Zakary L. and Derek D. Rucker (2007), "Attitude Certainty: A Review of Past Findings and Emerging Perspectives," *Social and Personality Psychology Compass*, 1 (1), 469–92.

Tormala, Zakary L. and Derek D. Rucker (2018), "Attitude Certainty: Antecedents, Consequences, and New Directions," *Consumer Psychology Review*, 1 (1), 72–89.

USA Today Ad Meter (2017), "2017 Results," (accessed April 1, 2021), https://admeter.usatoday.com/results/2017.

Vincze, Veronika, György Szarvas, Richárd Farkas, György Móra, and János Csirik (2008), "The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and Their Scopes," *BMC Bioinformatics*, 9 (11, Suppl.), S9.

Visser, Penny S. and Robert R. Mirabile (2004), "Attitudes in the Social Context: The Impact of Social Network Composition on Individual-Level Attitude Strength," *Journal of Personality and Social Psychology*, 87 (6), 779–95.

Web of Science (2021), "Sentiment Analysis," (accessed March 1, 2021), https://www.webofknowledge.com/.

Wicker, Allan W. (1969), "Attitudes Versus Actions: The Relationship of Verbal and Overt Behavioral Responses to Attitude Objects," *Journal of Social Issues*, 25 (4), 41–78.