Behav Res
DOI 10.3758/s13428-017-0975-6

The Evaluative Lexicon 2.0: The measurement of emotionality, extremity, and valence in language

Matthew D. Rocklage[1] · **Derek D. Rucker**[1] · **Loran F. Nordgren**[1]

Abstract The rapid expansion of the Internet and the availability of vast repositories of natural text provide researchers with the immense opportunity to study human reactions, opinions, and behavior on a massive scale. To help researchers take advantage of this new frontier, the present work introduces and validates the Evaluative Lexicon 2.0 (EL 2.0)—a quantitative linguistic tool that specializes in the measurement of the emotionality of individuals' evaluations in text. Specifically, the EL 2.0 utilizes natural language to measure the emotionality, extremity, and valence of evaluative reactions and attitudes. The present article describes how we used a combination of 9 million real-world online reviews and over 1,500 participant judges to construct the EL 2.0 and an additional 5.7 million reviews to validate it. To assess its unique value, the EL 2.0 is compared with two other prominent text analysis tools—LIWC and Warriner et al.'s (*Behavior Research Methods, 45,* 1191–1207, 2013) wordlist. The EL 2.0 is comparatively distinct in its ability to measure emotionality and explains a significantly greater proportion of the variance in individuals' evaluations. The EL 2.0 can be used with any data that involve speech or writing and provides researchers with the opportunity to capture evaluative reactions both in the laboratory and "in the wild." The EL 2.0 wordlist and normative emotionality, extremity, and valence ratings are freely available from www.evaluativelexicon.com.

Keywords Emotion · Text analysis · Language · Attitudes · Cognition

Over the last decade, the rapid expansion of the Internet and the availability of vast repositories of natural text have allowed researchers to systematically study individuals' reactions, opinions, and behavior on a massive scale. Due to this expansion, an immense opportunity now exists to study human responses and behavior "in the wild." As part of the effort to quantify responses in natural text, Rocklage and Fazio (2015) developed and introduced a novel computational tool called the Evaluative Lexicon (EL). The EL utilizes natural language to quantify the degree to which an individual's attitude or reaction is based on emotion (its *emotionality*), whether the reaction is positive or negative (its *valence*), and the extent of that positivity or negativity (its *extremity*).

The present research was undertaken with four primary objectives. First, we provide readers with a brief primer on the original EL. Second, we significantly expand the capabilities of the original EL. Specifically, we use an iterative, data-driven approach to catalog the words that are most indicative of individuals' evaluative reactions and opinions. As a result, we expand the wordlist of the original EL from 94 to 1,541 words in the EL 2.0—an increase of nearly 1600%. Third, we validate the EL as a measure of both individuals' opinions and their underlying emotionality using natural text. Finally, we differentiate the EL and its measurement of emotionality from two other text analysis tools popular within psychology: Linguistic Inquiry and Word Count (LIWC; Pennebaker,

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-017-0975-6) contains supplementary material, which is available to authorized users.

✉ Matthew D. Rocklage
matthew.rocklage@kellogg.northwestern.edu

[1] Kellogg School of Management, Northwestern University, 2211 Campus Drive, Evanston, Illinois 60208, USA

Published online: 19 October 2017

Boyd, Jordan, & Blackburn, 2015) and Warriner et al.'s wordlist (Warriner, Kuperman, & Brysbaert, 2013).

The Evaluative Lexicon

A particular focus of the EL is to provide insight into whether an individual has an evaluation that is more emotional in nature versus one that is less emotional and more cognitive. This distinction between emotion and cognition can be traced back at least as far as the writings of Aristotle in the 4th century BCE (see also Cicero, 1986), and it has received extensive study within the behavioral sciences (Haidt, 2001; Lazarus, 1982; Metcalfe & Mischel, 1999; Mischel & Shoda, 1995; Pham, 2007; Zajonc, 1980). This distinction has also been the subject of a great deal of work within the domain of attitudes. Attitude researchers have defined *emotionally-based* attitudes as evaluations based more on the feelings that a person has about an object, and *cognitively-based* attitudes as evaluations based more on a person's beliefs about the object and its properties (see Petty, Fabrigar, & Wegener, 2003, for a review). As a testament to the importance of this distinction for evaluations, research has shown that both emotional and cognitive reactions are integral to understanding individuals' overall evaluations (Abelson, Kinder, Peters, & Fiske, 1982; Bagozzi & Burnkrant, 1979; Haddock, Zanna, & Esses, 1993) and that each has important downstream consequences (Eagly, Mladinic, & Otto, 1994; Fabrigar & Petty, 1999; Huskinson & Haddock, 2004; Lavine, Thomsen, Zanna, & Borgida, 1998; Shiv & Fedorikhin, 1999; van den Berg, Manstead, van der Pligt, & Wigboldus, 2006; Williams & Drolet, 2005).

How does one know whether a person holds an attitude that is more emotional or cognitive in nature? One approach is to examine the language a person uses to convey that attitude. Indeed, we could easily convey our attitude toward something by simply saying that we "like" or "dislike" it, but instead we have a myriad of words to communicate our attitude, from "valuable" to "amazing" and from "objectionable" to "terrible." On the basis of this observation, Rocklage and Fazio (2015) developed the EL to use these differences in language to measure whether an individual has an attitude that is more emotional versus cognitive in nature.

To quantify individuals' language for its emotionality, Rocklage and Fazio (2015) generated a list of 94 adjectives—for example, "valuable," "fantastic," and "magnificent"—and obtained normative ratings from a set of judges with respect to the emotionality implied by each adjective, as well as its implied valence and extremity (the deviation from the midpoint of the valence scale). For instance, on the basis of these normative ratings, an individual who used the adjective "fantastic" would score a 6.64 out of 9.00 on emotionality, 8.57 out of 9.00 on positivity, and 4.07 out of 4.50 on extremity. Thus, this person would be expected to hold a quite positive and emotionally-based attitude. On the other hand, an individual who used the adjective "valuable" would score lower on emotionality (3.98) as well as on both positivity (7.68) and extremity (3.18). This person would also be expected to hold a positive attitude, but one based less on emotion and not as extreme. Thus, each time an individual uses one of the EL adjectives, these values can be imputed and the adjectives can be numerically represented. These normative values, as will be discussed, can be used to predict outcomes of interest.

By quantifying both the emotionality and extremity of the adjectives people use, Rocklage and Fazio (2015, 2016) found that more emotional reactions tend to be more extreme in their positivity or negativity. This natural association can be seen in the adjective "fantastic," which is both positive and emotional, and the adjective "valuable," which is relatively less positive as well as less emotional. Despite this relationship, the emotionality and extremity of individuals' evaluations are distinct. Take, for example, the words "wonderful" and "outstanding." Although they imply equally extreme positive evaluations (both imply approximately a 4.00 on extremity), they differ in their implied emotionality (6.98 and 5.92, respectively). "Wonderful" signals an evaluation based relatively more on emotion. Moreover, differences in extremity and emotionality account for unique variance when predicting outcomes of interest (Rocklage & Fazio, 2015, 2016). Thus, a strength of the EL is that it can provide measures of both the emotionality and extremity of individuals' attitudes, even with a single word.

The Evaluative Lexicon: validation and predictive ability

Rocklage and Fazio (2015) validated the EL's ability to measure emotionality using both laboratory-based experiments and archival text. In the laboratory-based research (Study 2), they experimentally created positive or negative attitudes toward a fictitious aquatic animal that were either emotionally- or cognitively-based. In the positive *emotion* condition, for example, they provided participants with a narrative in which a swimmer rode on the back of the animal through the water. In the positive *cognitive* condition, they provided participants with an encyclopedic entry that indicated the animal, for example, provided essential nutrients for coastal communities (see also Crites, Fabrigar, & Petty, 1994). On the basis of the adjectives individuals used to describe their evaluations, the EL successfully predicted participants' condition (e.g., emotional or cognitive) 88.2% of the time. As further evidence of the ability of the EL to capture both emotional and cognitive responses, participants used more emotional adjectives in the emotional condition (e.g., "amazing," "wonderful," and "delightful") and more cognitive, unemotional adjectives in the cognitive condition (e.g., "helpful" "beneficial," and "valuable").

Rocklage and Fazio (2015, Study 3) validated the EL in natural text using a large number of real-world Amazon.com product reviews. The authors found that the emotionality of an adjective predicted the presence of other words that signaled individuals had a more emotional versus cognitive evaluation. Specifically, the more emotional an adjective, the more often it occurred alongside the verb "feel" within the reviews and, conversely, the more cognitive and unemotional an adjective the more it was accompanied by the verb "believe." This association was specific to the emotionality of the adjectives as their implied extremity was not related to the use of these verbs.

The EL can also predict individuals' judgments. Across 5.9 million product reviews, the greater emotionality (versus cognition) reviewers expressed in their text, the more extreme their final star ratings (Rocklage & Fazio, 2015, Study 3). These results were extended into the laboratory, where greater emotionality was a better predictor of individuals' final judgments when they were required to come to a quick, dichotomous decision (Rocklage & Fazio, 2016). Taken together, these results demonstrate both the ability of the EL to measure the emotionality of individuals' reactions as well as the importance of studying emotionality via people's language.

Expansion of the Evaluative Lexicon: an overview

The original EL we have overviewed—hereafter referred to as the EL 1.0—was optimized to be comprehensive in its coverage of a wide range of emotionality, extremity, and valence. However, it was not designed with the goal of creating an expansive list of words. Indeed, the EL 1.0 consists of 94 adjectives, a size which can limit its application in natural text. The key objective of the present work was to dramatically increase the size and scope of the EL and to then validate the expanded wordlist as a measure of individuals' attitudes and their emotionality. As a secondary objective we also sought to differentiate the EL 2.0 from other tools available for text analysis. Before providing concrete details of each step, we first overview the general approach used in the present research.

As a first step to increase the size of the EL, we extracted the most frequently used words from five different real-world sources. From this list, we targeted those words that had the potential to indicate an evaluative response as this is the central purpose of the EL. For example, whereas the word "amazing" is very likely to indicate an evaluation, a word such as "dog," though potentially positive, is unlikely to be indicative of an individual's evaluation. To that end, we refined the wordlist through multiple rounds of ratings from a large set of trained judges to assess the extent to which each word was likely to imply an evaluation. Next, we used those words judged as likely to signal an evaluation as seeds to automatically propagate additional evaluative synonyms. Finally, we used a data-driven approach to investigate whether each word is used consistently across a wide range of topics in real-world contexts. We retained only those words likely to imply an evaluation and one that is relatively consistent across topics.

After developing the wordlist that comprises the EL 2.0, we validated the EL 2.0 by conceptually replicating previous findings. Specially, we demonstrated that more emotional EL words are more likely to be accompanied by words and phrases that signal a more emotional reaction (e.g., "I feel"), whereas more cognitive EL words are more likely to be accompanied by words and phrases related to cognition (e.g., "I believe"). We then replicated the effect reported by Rocklage and Fazio (2015), whereby greater implied emotionality (vs. cognition) tends to predict more extreme summary judgments.

Next, we compared the EL 2.0 with two prominent approaches in psychology used to quantify language: LIWC (Pennebaker et al., 2015) and Warriner et al.'s (2013) wordlist. Although these wordlists have a variety of valuable functions, each also purports to measure an aspect of emotionality. To date, no version of the EL has been compared to these linguistic approaches. As such, although the EL was explicitly constructed to capture emotionality, it is unclear whether it offers unique capabilities relative to these existing approaches. Thus, to investigate whether the EL 2.0 has utility above and beyond existing tools, we compared it to each of these approaches.

Creation of the Evaluative Lexicon 2.0

Acquisition of the initial wordlist

To create a wordlist with the ability to measure individuals' reactions and evaluations across a wide range of topics, we first compiled the most frequently used words across five large, diverse sources.

The first source was Amazon.com. We obtained 5.9 million product reviews written on the Amazon website from 1996 to 2006 (Jindal & Liu, 2008). These reviews contained individuals' evaluative reactions to 1.2 million products, ranging from vacuum cleaners and toasters to music, books, and movies. The second source was TripAdvisor.com, from which we obtained 1.6 million reviews across 12,746 different hotels from 2001 to 2012 (Wang, Lu, & Zhai, 2011). These hotels ranged from luxury hotels to budget motels. The third source was Yelp.com, from which we obtained 1.6 million reviews across 60,785 businesses from 2004 to 2015. These businesses ranged from restaurants to clothing stores to health services. Finally, we also obtained 9,388 U. S. movie and TV show scripts from 1900 to 2007 (Brysbaert & New, 2009) and 1 million tweets extracted from Twitter.com's streaming application program interface (API) over a

24-h period. Taken together, these five sources resulted in 1.5 billion words in total, and 6.2 million unique words.[1]

From each of these five sources we extracted the 10,000 most frequently used words that were not Web addresses, Twitter handles, Twitter hashtags, or common "stop words" (e.g., "a," "an," "the"; Bird, Klein, & Loper, 2009). We then combined the 10,000 most frequently used words from each source and removed any duplicate words, but we included similar words with different tenses (e.g., "love," "loved"). We were thus left with 21,189 unique words.[2]

Refinement of the wordlist, part 1: judgments of the evaluative nature of each word

To refine the wordlist, we utilized a three-step procedure. To begin, we elicited two sets of judgments from participants to assess the extent to which the words we extracted provided an indication of individuals' evaluations. This set of judgments was used to separate the words into three categories: words unlikely, somewhat likely, and likely to signal an evaluation (Step 1). We then removed those words unlikely to signal an evaluation from the wordlist and, to enhance the coverage of the wordlist, used seed-word propagation to generate synonyms for words judged as likely to signal an evaluation (Step 2). The words that were assessed as somewhat likely to signal an evaluation, as well as the synonyms just mentioned, were submitted to a second set of judges who assessed the extent to which they signal an evaluation (Step 3). The specific details of each of these steps follow.

Step 1: Initial word assessment We asked 716 MTurk participants to rate approximately 300 words each (213,690 total judgments) on the extent to which each word tended to imply an evaluation (1 = *almost never*, 5 = *almost always*). We also allowed participants to indicate that they did not know a word, since their judgments would likely be haphazard or inaccurate in such cases. As with all of the rating procedures in this article, participants were paid $1.00 for their time.

Before participants proceeded to the judgment task, we presented them with four words ("big," "hot," "anger," and "excellent") to ensure that they understood the instructions and to provide training. This training procedure allowed participants to gain an understanding of the latent space as well as provided them with a means to judge subsequent words. Participants were first asked to rate the extent to which the practice word implied an evaluation. They were then provided with an explanation as to why that word might or might not signal an evaluation. For instance, after participants judged the word "big," they were informed that although "big" can imply an evaluation in some contexts, it may not imply an evaluation in other contexts. Participants were then provided with the example that whereas a review that describes a television screen as "big" could signal a positive evaluation, it is ambiguous whether a review that describes a chair as "big" is evaluative or simply descriptive of its size. Thus, they should rate "big" lower on the scale. To further reinforce this point, they were then given the example of the word "amazing" and how this word would be considered clearly evaluative across various topics, and thus they should issue it a high rating.

After participants read the instructions and responded to the four practice words, they judged 300 randomly selected words from the 21,189 words we had extracted previously. We then asked participants whether or not their first language was English. We collected judgments until each word had been judged by approximately ten participants, on average.

To refine the wordlist, we retained participants only if they indicated that their first language was English. This led to a sample of 660 participants. Next, we removed those words that were judged, on average, as "almost never" indicative of an evaluation (i.e., words that received a score lower than 2; 14,500 words) as well as those that over half of the participants did not know (2,191 words).

Step 2: Seed-word propagation After we obtained the evaluations of the words in Step 1, we expanded the list via seed-word propagation. Specifically, we identified a set of seed words and used these seeds (e.g., "fantastic") to propagate additional candidate words (e.g., "wondrous" and "tremendous"). We identified seed words as those words that were judged as either "often" or "almost always" (received a score of 4 or greater) indicative of an evaluation (1,330 words), as well as the 94 adjectives from the EL 1.0 (1,359 unique seed words). Using synsets from WordNet (Miller, 1995), we generated 5,551 synonyms from these seeds.

Step 3: Further word assessment In the third step of refining the wordlist, we asked a new set of MTurk participants to judge the extent to which the synonyms generated in Step 2 indicated an evaluation. In addition, participants judged the words from Step 1 that had been judged as "seldom" or only "occasionally" indicative of an evaluation (i.e., that received a score of 2 and greater, but lower than 4; 3,168 words). The latter words were included as an additional means to determine whether or not these "borderline" cases should be taken

[1] These 6.2 million unique words also included, for example, misspellings, names of musicians and writers, product and brand names, Web addresses, and Twitter handles and hashtags. These items were filtered in subsequent stages.

[2] A notable limitation of focusing on the most frequently used words is that such an approach does not capture every word in the English language. However, focusing on the most frequent words allowed for a structured approach that both was feasible to implement and served the practical goal of enhancing the likelihood that the EL would properly capture individuals' evaluative reactions in a given piece of text.

as indicative of an evaluation. We removed duplicate words as well as those that had been eliminated (those that received scores lower than 2) or retained (scores greater than or equal to 4) in Step 1. This left 6,264 words for judgment.

We asked 210 MTurk participants to judge approximately 300 words each (67,712 total judgments). In addition to the words individuals had used for training in Step 1, we asked them to train with four additional words ("definitely," "smart," "despise," and "useful"). As before, we provided participants with an explanation for why each word might or might not signal an evaluation. We used the same rating scale as in Step 1. After participants completed their judgments, we asked them to report whether English was their first language.

We collected responses until each word had been judged, on average, approximately ten times. As before, we retained those judgments that came from participants whose first language was English (participants: $n = 183$). We then removed those words that participants had judged as "almost never," "seldom," or "occasionally" indicative of an evaluation (i.e., that received a score lower than 4; 4,373 words) and those that more than half of the participants did not know (277 words). We were left with 1,614 words from this second step. After combining these words with the previous set of words retained in Step 1, we were left with a total of 2,973 total unique words.

Quantification of each word's implied emotionality, extremity, and valence

We next used two large samples of participants to judge the valence and emotionality of the 2,973 unique words from our previous steps. These ratings would serve as the basis for the Evaluative Lexicon 2.0 and would be imputed each time an individual used one of the words from this list.

We followed Rocklage and Fazio's (2015) procedure to elicit the emotionality and valence ratings. Specifically, participants were provided with the following instructions:

> Sometimes when we evaluate an object, person, or event, we do so on the basis of an emotional reaction to the object, person, or event. That is, our emotions determine whether we conclude that we like or dislike the object, person, or event. Indeed, some evaluative terms appear to imply that the evaluation was arrived at on the basis of emotion. Using the scale on the next page, please rate the extent to which each term implies an evaluation based on emotion.

Participants were then given a 0 to 9 scale on which 0 indicated that the word was *not at all emotional* and 9 that the word was *very emotional*. They then rated their randomly selected set of words.

A separate set of participants rated the valence implied by each word and were provided the following instructions:

> When we evaluate an object, person, or event, we often use terms such as those listed on the next page. Some evaluative terms imply a negative evaluation and some a positive evaluation. Using the scale on the next page, please rate the evaluation implied by each term.

These participants were given a 0 to 9 scale on which 0 indicated that the word was *very negative* and 9 that the word was *very positive*. We also provided both sets of judges the ability to indicate they did not know a word.

Participants judged approximately 270 randomly chosen words from the 2,973 words we had obtained. In addition, we selected 30 words from the EL 1.0 that covered the possible range of valence present from that list, as well as a separate set of 30 words that covered the possible range of emotionality. Participants who judged the valence of the words received a set of 30 EL 1.0 adjectives that spanned the range of valence, whereas participants who judged emotionality received a set of 30 EL 1.0 adjectives that spanned the range of emotionality. As will be detailed subsequently, we included these words to ensure quality responses. Thus, in total, we asked each participant to judge approximately 300 words. We then asked participants whether or not their native language was English.

After we obtained an initial set of ratings, we took additional steps to ensure the quality of the word judgments. First, we retained only those participants whose native language was English. Second, we calculated a correlation between each participant's ratings of the 30 EL 1.0 words and the normative ratings provided by the EL 1.0 judges. Participants were retained if their judgments of these 30 words were significantly correlated ($p \leq .05$) with the EL 1.0 normative ratings. In essence, if participants' responses to these words were significantly correlated with the EL 1.0 normative ratings, this provided evidence that they understood and were engaged with the present rating task. One hundred ten participants were unable to be included on the basis of these criteria. We then elicited additional judgments from new participants that matched these criteria until we had, on average, approximately 30 unique ratings per word. In all, 305 participants made a total of 88,705 judgments for valence, and 334 participants made a total of 96,912 judgments for emotionality. Finally, we removed words that were common misspellings or that fewer than half of the participants knew (137 words). We were left with 2,836 total words.

To quantify the normative valence and emotionality implied by each word, we averaged each word's valence ratings across participants and then averaged each word's emotionality ratings. Following the original procedures by Rocklage and Fazio (2015), we then quantified the extremity implied by each word as the distance of each participant's valence rating from the midpoint of the valence scale (i.e., the absolute

value of valence minus 4.50). Any deviation from this midpoint indicates a more extreme rating. We then averaged together these extremity ratings across participants.

Refinement of the wordlist, part 2: examination of each word's use across topics

Given that our aim was to provide a set of words sufficiently general to be used across topics, we sought to retain only those words that were consistently associated with either positivity or negativity. For instance, whereas the word "fair" can be used to describe a person positively, it can also be used to signal a rather lukewarm endorsement of food. The word "wonderful," on the other hand, would be quite consistent in signaling a positive evaluation across topics. To this end, we assessed whether the implied valence of each word was relatively consistent with respect to how each word is used across topics in real-world contexts.

We returned to those online reviews used to generate the wordlist. For each review, users provided both their reaction in written text and a final summary judgment in the form of a one- to five-star rating. Of importance, the star rating allowed us to assess the degree to which the language they used was related to generally positive versus negative reactions. As such, these ratings provided a means of testing the relative consistency of the valence implied by each word.

To begin, we coded each of the 9 million reviews from Amazon, TripAdvisor, and Yelp for whether or not each word was included in the review. As we detailed previously, these reviews covered a wide range of 1.27 million products and businesses, and should therefore provide a reasonable indication of the extent to which a word is used relatively consistently across topics. If the word was used in the review, that review received a "1" for that particular word; otherwise the word received a "0" for that review. For this analysis and all the subsequent analyses that involved the EL, we did not count a word within a review if it was preceded by a negation (e.g., "was not," "is not," etc.). It is not clear, for example, that evaluating an object as "not amazing" should be treated as the opposite of "amazing."

We then coded each review for whether the reviewers indicated they were generally positive (four- and five-star reviews; coded as "1") or negative (one- and two-star reviews; coded as "0") toward the product, hotel, or business they wrote about. In essence, if the word was more likely to be present in positive (vs. negative) reviews and was rated as implying a positive reaction by the judges (above the 4.50 midpoint of the valence scale), this offered evidence that the word is used relatively consistently to indicate a positive reaction across topics.

For each word and each of the three review sources, we used logistic regression to predict the probability that a word would be associated with a positive versus a negative review. We retained a word if it was more likely to be associated with positive (or with negative) reviews (i.e., if the sign of the coefficient was consistently in either the positive or the negative direction) in each of the sources in which it was present, and if it was judged to be positive (negative) by the external raters.[3] This approach was used in an effort to provide a conservative estimate of the extent to which the word implied similar evaluations across topics. On the basis of these criteria, we were left with the 1,541 words that constitute the Evaluative Lexicon 2.0 (see Table 1 for a summary of the number of words added and removed at each stage).

Details of the Evaluative Lexicon 2.0

We first explored the properties of the EL 2.0 to provide initial validation. To begin, we calculated a measure of the reliability of the ratings. Following previous research for measuring reliability (Stadthagen-Gonzalez, Imbault, Pérez Sánchez, & Brysbaert, 2017), we randomly selected half of the participants and then calculated the average emotionality or valence for each of the 1,541 words 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, we were left with 100 samples that contained the average emotionality or valence for each of the 1,541 words, based on a randomly selected subset of participants. We then correlated the 100 samples with each of the others (4,950 possible pairings). As evidence of their consistency, strong correlations emerged across the samples for both emotionality (r_{avg} = .910; 95% confidence interval [CI]: [.909, .911]) and valence (r_{avg} = .988; 95% CI: [.987, .988]).

To further assess the validity of the ratings, we correlated the average ratings obtained from the present set of judges with the original normative ratings of those words from the original set of judges from the EL 1.0. The valence [$r(64) = .99$, $p < .001$], extremity [$r(64) = .87$, $p < .001$], and emotionality [$r(64) = .89$, $p < .001$] ratings were extremely consistent between the two sets of ratings. These results indicate that the ratings showed good consistency even across time and with different participants.[4] The average valence implied by the words was 4.17 out

[3] Due to the relatively low baserates of certain words (e.g., "condemnable"), the logistic regression models were unable to converge for some of the review sources. In these cases, we relied on those sources for which the regression models did converge. For instance, if the regression model did not converge for one of the sources, but the word was used consistently across the other two sources and was judged similarly by the external raters, then the word was retained. If the word was not present across any of the sources, then it was not included in the final wordlist (31 total words; e.g., "aristocratical," "bedaze," and "thriftlessness").

[4] Although a majority of the EL 1.0 words were also included in the EL 2.0 (66), some words were not included. In particular, the explicit goal of the EL 2.0 to contain words that were consistent across topics led to the removal of a portion of the words from the EL 1.0. For example, the word "con" was not included in the EL 2.0 wordlist because this word often signals a negative reaction, but it can also be used in a non-evaluative manner—for example, to reference popular dishes at a Mexican restaurant such as "chile con queso." Thus, "con" did not make the EL 2.0 wordlist due its different uses across topics.

Table 1 Summary of word generation and selection for the Evaluative Lexicon 2.0 (EL 2.0).

	Initial Number of Words	Number of Words Added/Removed	Details of Addition/Removal
Initial wordlist	6.2 million	– 6.18 million	Extracted the 10,000 most frequently used words from each of the five sources
Refinement of the wordlist, Part 1			
Initial word assessment (Step 1)	21,189	– 16,691	Removed words unknown by a majority of judges and those judged unlikely to be indicative of an evaluative reaction
Seed-word propagation (Step 2)	1,359	+ 1,766	Propagated synonyms for those words judged likely to be indicative of an evaluative reaction. Added unique synonyms to main wordlist
Further word assessment (Step 3)	6,264	– 3,291	Removed words unknown by a majority of judges and those judged unlikely to be indicative of an evaluative reaction
Quantification of each word	2,973	– 137	Removed words unknown by a majority of judges
Refinement of the wordlist, Part 2	2,836	– 1,295	Removed words not used consistently across topics
EL 2.0	1,541		

Numbers in the "Initial Number of Words" column represent the number of starting words for that step and not necessarily the overall number of words obtained up to that step.

of 9.00 (SD = 2.67), the average extremity was 2.56 out of 4.50 (SD = 0.83), and the average emotionality was 4.42 out of 9.00 (SD = 1.51).

The correlations between the different dimensions of the expanded wordlist were also similar to those obtained from the original normative ratings reported previously (Rocklage & Fazio, 2015). Specifically, the valence and extremity [$r(1539) = -.10$, $p < .001$] as well as the valence and emotionality [$r(1539) = -.13$, $p < .001$] ratings showed little association, though negative words tended to be slightly more extreme and emotional. Most importantly, as a result of the tendency for emotional reactions to be more extreme, and in line with the EL 1.0 normative ratings, the more emotional a word was, the more extreme it was [$r(1539) = .47$, $p < .001$]. However, the modest size of this correlation demonstrates that emotionality and extremity are related but separable constructs.[5]

Taken together, these correlations demonstrate that the expanded wordlist is in line with past work, in terms of both the ratings obtained and the associations between the different dimensions. However, whereas for the EL 1.0 we identified 94 adjectives, the EL 2.0 consists of 1,541 words that cover different parts of speech—an increase of nearly 1600%.

Validation of the Evaluative Lexicon 2.0

Validation of emotionality

Given that a key focus of the EL is to measure the emotionality of individuals' evaluations, we sought to demonstrate that the EL 2.0 conceptually replicates prior results. Specifically, to validate the EL 2.0 we tested whether more emotional words from the wordlist would accompany more emotional reactions (e.g., "feel") versus more unemotional, cognitive reactions (e.g., "believe"; Rocklage & Fazio, 2015) in real-world text.

To this end, we obtained four new sets of online reviews (see McAuley, Pandey, & Leskovec, 2015) that were not utilized to generate the EL 2.0 wordlist. We selected categories that we believed would differ systematically as this provided a means to assess the extent to which our results held across a wide assortment of different topics and product types. Specifically, we used four distinct categories of products written between 2007 and 2014 on Amazon.com that ranged from those we hypothesized may naturally evoke emotionality—those more hedonic in nature—to those we hypothesized would be less likely to evoke emotionality—those more utilitarian in nature (Batra & Ahtola, 1991; Hirschman & Holbrook, 1982; Voss,

[5] For interested readers, this correlation also indicates a curvilinear relationship between valence and emotionality. Indeed, using regression, a squared valence term (standardized) was significant when predicting the emotionality of a word [$B = 1.33$, $t(1538) = 21.55$, $p < .001$]. This term indicates that emotionality increases as valence becomes extremely positive or negative.

In the present study we did not ask participants for additional demographic information, and thus cannot investigate demographic differences. However, results from the EL 1.0 indicate that there is high agreement, for example, between male and female individuals' judgments [valence: $r(92) = .99$, $p < .001$; emotionality: $r(92) = .92$, $p < .001$; Rocklage & Fazio, 2015]. We anticipate that these results would be similar for the EL 2.0, given the other similarities with the EL 1.0 that we have demonstrated here. Nevertheless, future work might benefit from more of an examination of whether there are meaningful differences across demographics.

Spangenberg, & Grohmann, 2003). Those categories we anticipated to naturally evoke greater emotionality were toys and games (2,159,996 reviews, 1,285,926 reviewers, 322,278 products) and music instruments and gear (486,067 reviews, 329,224 reviewers, 80,883 products). Those categories we anticipated to evoke less emotionality were office materials (1,206,446 reviews, 882,181 reviewers, 128,533 products) and tools and home improvement merchandise (1,888,208 reviews, 1,285,926 reviewers, 258,640 products). Across the four product categories we obtained 5.7 million reviews (words per review: M = 68.77), 3.1 million unique reviewers, and 790,334 different products.

To examine whether the categories did indeed differ in the extent to which they naturally elicited emotionality, we calculated a weighted average of the implied emotionality for each product category using the normative EL 2.0 ratings we had obtained for those reviews. To illustrate, consider this sentence: "This game was amazing—you really get into the action. The storyline was also amazing and its whole execution was flawless." "Amazing" has an emotionality score of 6.60, whereas "flawless" has an emotionality score of 3.05. Thus, we would quantify the emotionality implied by this reviewer as (2 * 6.60 + 1 * 3.05)/3 = 5.42. We then averaged the emotionality of the reviews from each category. In support of our a priori categorization, the toys and games category elicited the greatest degree of emotionality from reviewers, as measured by the EL 2.0 (M = 4.98, SD = 1.44), followed by music instruments and gear (M = 4.30, SD = 1.24), office materials (M = 4.14, SD = 1.31), and finally the tools and home improvement category (M = 4.04, SD = 1.24). Each of these categories was statistically significantly different from one another (ps < .001). These averages provide additional evidence in support of the appropriateness of the EL 2.0 as a measure of emotionality.

To further validate the normative emotionality ratings, we constructed a list of phrases that we anticipated would accompany more emotional reactions versus those that would accompany relatively more cognitive, unemotional reactions within the reviews. The phrases that were likely to signal more emotional reactions were "I felt," "I feel," "feelings," "emotion," "emotions," and "emotional." The phrases that were likely to signal more cognitive, unemotional reactions were "I believed," "I believe," "I considered," and "I consider."[6] For those reviews that contained at least one of those phrases, we counted the number of times an EL 2.0 word was used in the same review as an emotion-signaling phrase versus a cognition-signaling phrase. In other words, we included a review only if it contained either an emotion- *or* a cognition-signaling phrase, but did not include the review if it contained both types of phrases. A total of 81,917 reviews used an emotion- versus a cognition-signaling phrase (31,773 and 50,144 reviews, respectively). Of this total, 78,897 reviews also used an EL 2.0 word (96% of the total; 98% of the emotion-signaling reviews and 93% of the cognition-signaling reviews used an EL word). Following past research (Rocklage & Fazio, 2015), we then calculated the proportion of times each EL 2.0 word was used with the emotion-signaling phrases versus the total times it was used with either the emotion- or the cognition-signaling phrases [times used with emotion-signaling phrases/(times used with emotion-signaling phrases + times used with cognition-signaling phrases)]. This proportion indexes the extent to which to an EL word co-occurs with an emotion-signaling phrase, whereas 1 minus this proportion indexes the extent to which it co-occurs with a cognition-signaling phrase.

To construct the final list of words for the present analyses, words that were not used in a review that contained either of these sets of phrases were not included (n = 103). We also did not include the word "feel" from the EL 2.0 wordlist given that "feel" is one of the very words we had selected to validate the EL 2.0. Finally, so that relatively rare words would not unduly influence the final results, we also did not include those words that were used fewer than ten times (n = 330). We were thus left with 1,107 words for analysis.

We correlated the emotionality proportion (M = .67, SD = .11) with the normative emotionality ratings we had just obtained. We controlled for the extremity of the word in order to assess emotionality per se. As hypothesized, we found that the greater the emotionality a word implied, the more it was used alongside phrases that signaled the presence of an emotional reaction [r(1104) = .26, p < .001]. This correlation also indicates that the EL 2.0 words that implied a more cognitively-based reaction were used relatively more with phrases that signaled cognition.

To assess whether these results were specific to the emotionality of the words, we next correlated the implied extremity of each word with the emotionality proportion. Controlling for the emotionality of each word, the extremity of the words was not related to the emotionality proportion [r(1104) = – .04, p = .22]. This result indicates that extremity reflects a separable dimension of individuals' evaluative reactions apart from emotionality. Moreover, this outcome provides evidence that the emotion- and cognition-signaling phrases do not simply measure the relative tentativeness of

[6] For interested readers, we also considered including the terms "thought" and "think" as part of the cognition-signaling words. However, a more detailed analysis of the reviews that utilized these words revealed that, in fact, they also appear to signal "hedging" on the part of the reviewer thereby making these words less clear signals of cognition. Indeed, close synonyms of "think" are "guess" and "suppose," both of which signal uncertainty. As evidence of this, reviewers who used the words "think" and "thought" issued less extreme final star ratings (M = 1.44, SD = 0.71)—calculated as the absolute value of the deviation from the midpoint of the star rating scale—than did reviewers who used the other cognition-signaling phrases (M = 1.52, SD = 0.67) [t(515353) = 18.44, p < .001]. Nevertheless, even when including "think" and "thought" in those analyses reported in the main text, the results were similar.

individuals' reactions and thereby reinforces these phrases as assessing emotionality per se. Taken together, these results demonstrate that the emotionality dimension of the EL 2.0 appears valid, given its co-occurrence with phrases signaling emotional reactions (vs. cognition) in real-world text.

Prediction of self-reported evaluations

To further validate the wordlist, we sought to replicate the EL 1.0's ability to predict individuals' self-reported evaluations. Rocklage and Fazio (2015) found that greater implied emotionality, extremity, and the number of EL 1.0 words predicted more extreme positive or negative evaluations as measured via the final star ratings individuals issued a product (out of a possible five stars). Given our wide range of product categories, we had the additional opportunity to examine whether these relations held across products that differed naturally in the emotionality they elicit.

For each review, we calculated a weighted average of the implied emotionality and extremity for the positive and negative words separately (i.e., those with a normative valence rating of 4.50 and above or those with a rating below 4.50, respectively). To illustrate, take the following sentences as examples of quantifying implied emotionality: "This wrench was great. It has its flaws and it can be cumbersome, but it was great." We would quantify the emotionality implied by the positive words ("great"): (2 * 4.74)/2 = 4.74. Then we would quantify the emotionality implied by the negative words ("flaws" and "cumbersome"): (1 * 3.04 + 1 * 3.83)/2 = 3.44. We then subtract these scores from each other to quantify the extent to which individuals are relatively more emotional in their positive versus negative reactions: 4.74 – 3.44 = 1.30. Thus, this reviewer would be assessed as relatively more emotional in their positive reactions. We followed this same approach to calculate the implied extremity of individuals' reactions. Finally, we calculated the number of positive versus negative words each reviewer used (number of positive minus negative words). Given that this term quantifies the extent to which reviewers focus on their positive versus negative reactions, this variable has been termed *valence focus* (Rocklage & Fazio, 2015, 2016). After retaining those reviews that were covered by the EL 2.0, we were left with 87.2% of the sample (5,005,041 of 5,740,717 reviews; EL words per review: $M = 3.01$).

As we do for all subsequent regression analyses, we standardized each term before entering it into the regression model. We then used the standardized extremity, valence focus, and emotionality variables to predict the final star rating each reviewer provided. We replicated the results from the EL 1.0 (Rocklage & Fazio, 2015) with the EL 2.0: greater positive (negative) extremity [$B = 0.53$, $t(5005037) = 473.69$, $p < .001$], valence focus [$B = 0.30$, $t(5005037) = 593.56$, $p < .001$], and emotionality [$B = 0.04$, $t(5005037) = 32.25$, $p < .001$] all predicted more positive (vs. negative) star ratings. Moreover, these results were consistent across the four diverse categories, despite the differences in these categories' natural propensities to elicit emotionality (see Table 2).

Although extremity and valence focus were strong predictors of individuals' final judgments, emotionality appeared to be relatively less powerful. This finding is somewhat surprising given the greater predictive ability of emotionality in other settings (e.g., Rocklage & Fazio, 2016). However, one reason for this weaker effect might be the format of online reviews and the expectations present for reviewers. Specifically, when reviewers provide their evaluations, they are often expected to provide additional reasoning for why they had the positive or negative reactions they did in order to be helpful to others. This expectation may often lead them to provide more detail about the product and what elicited their reaction, all of which can be rather cognitive and unemotional. For example, reviewers might say they "loved" a product, but then feel compelled to explain that this was because it was "quiet," "sturdy," or "efficient." Thus, when averaging across the different words to obtain a measure of reviewers' emotionality, we may lose important information regarding the extent of their emotional reactions. As such, it is possible that the most emotional positive or negative reaction would be more indicative of the extent of the emotionality felt toward the product.

To examine the predictive power of people's most emotional reaction, we used the same regression model as above, but using the reviewers' most emotional reaction (most emotional positive minus most emotional negative word) in place of their average emotionality. This approach revealed much stronger effects of emotionality: The more emotional individuals' most emotional positive (negative) reaction, the more positive (negative) their final judgments [$B = 0.14$, $t(50005037) = 140.89$, $p < .001$]. These results held over and above extremity [$B = 0.45$, $t(50005037) = 474.71$, $p < .001$] and valence focus [$B = 0.27$, $t(50005037) = 499.06$, $p < .001$], and were consistent across the different product categories (see Table 2).[7]

[7] We also conducted these analyses using the positive and negative variables separately within a regression model. These results replicated those in the main text. As we hypothesized, positive [$B = 0.27$, $t(50005034) = 380.53$, $p < .001$] and negative [$B = -0.24$, $t(50005034) = 202.75$, $p < .001$] extremity, positive [$B = 0.08$, $t(50005034) = 110.49$, $p < .001$] and negative [$B = -0.03$, $t(50005034) = 26.62$, $p < .001$] emotionality, and positive [$B = 0.25$, $t(50005034) = 466.50$, $p < .001$] and negative [$B = -0.27$, $t(50005034) = 387.40$, $p < .001$] valence focus all predicted the star ratings in the expected direction.

Table 2 Standardized regression coefficients predicting reviewers' star ratings as a function of their average ($Emotionality_{Avg}$) and maximal ($Emotionality_{Max}$) expressed emotionality as assessed via the Evaluative Lexicon 2.0.

Predictor	*B*	*t*	Star Rating Change	Predictor	*B*	*t*	Star Rating Change
Toys/Games				Toys/Games			
Extremity	0.51	280.68	2.29	Extremity	0.42	271.44	1.88
Valence Focus	0.31	398.55	0.94	Valence Focus	0.27	322.87	0.81
$Emotionality_{Avg}$	0.08	50.33	0.42	$Emotionality_{Max}$	0.21	129.40	0.90
Music Gear				Music Gear			
Extremity	0.50	140.82	2.19	Extremity	0.39	130.12	1.72
Valence Focus	0.20	170.77	0.69	Valence Focus	0.17	141.08	0.60
$Emotionality_{Avg}$	0.05	14.44	0.26	$Emotionality_{Max}$	0.18	59.24	0.59
Office Materials				Office Materials			
Extremity	0.55	215.79	2.41	Extremity	0.47	224.72	2.09
Valence Focus	0.35	277.08	0.91	Valence Focus	0.31	229.60	0.82
$Emotionality_{Avg}$	0.04	15.51	0.21	$Emotionality_{Max}$	0.14	60.76	0.64
Home Improvement				Home Improvement			
Extremity	0.48	237.11	2.10	Extremity	0.41	246.26	1.83
Valence Focus	0.32	318.36	0.85	Valence Focus	0.30	267.61	0.78
$Emotionality_{Avg}$	0.02	11.61	0.13	$Emotionality_{Max}$	0.11	59.80	0.51

All coefficients are significant at $p < .001$. The "Star Rating Change" column indexes the difference in star rating between the minimum and maximum values possible in each product category for emotionality and extremity and from the 5th to 95th percentile for valence focus.

The star ratings that individuals issue are important in light of the direct impact that such ratings have on consumer purchases and business revenue (Chevalier & Mayzlin, 2006; Luca, 2011). To this end, following Rocklage and Fazio (2015), Table 2 also provides a more concrete measure of the impact of each dimension in terms of the final star ratings. Specifically, we calculated the change in star ratings from the reviews that were most negative to those most positive for the emotionality and extremity dimensions. Given that a small number of reviews in each category used a very large number of either positive or negative EL words, we used the values at the 5th and 95th percentiles for the valence focus variable.[8] To provide an example for emotionality from the toys-and-games category, above and beyond the extremity and valence focus variables, the difference between reviews with the greatest negative emotionality and reviews with the greatest positive emotionality corresponded to a change of approximately 1/2 star when using the average emotionality variable ($Emotionality_{Avg}$), and one star when using the maximal emotionality variable ($Emotionality_{Max}$). Notably, these changes in star ratings occurred above and beyond both the valence implied in each review and the number of positive versus negative words in that review, each of which has very obvious implications for individuals' final judgments. By controlling for these variables within the regression models and isolating the effect of the emotional versus cognitive words, we were in essence comparing the predictive ability of words such as "amazing" versus "perfection"—which are both quite positive (both 3.47 out of 4.50 on extremity) but differ greatly in their implied emotionality (6.60 and 4.27 out 9.00, respectively).

[8] Given the relative rarity of reviews that used a very large number of either positive or negative EL words in conjunction with the large number of reviews that we utilized, these reviews had little effect on the regression outcomes. Indeed, the results were very similar when we conducted the analyses without these reviews.

Differentiation of the Evaluative Lexicon 2.0 from other approaches

Given the validation of the EL 2.0, we next sought to evaluate its utility as compared to two other popular tools in psychology that are used to quantify language: Warriner et al.'s (2013) wordlist (referred to as "WRW" for simplicity) and Linguistic Inquiry and Word Count 2015 (LIWC; Pennebaker et al., 2015). We briefly overview these wordlists and the approaches they use to quantify text.

WRW The WRW is a large corpus of around 14,000 words collected from existing databases and has been put forth as a tool to estimate individuals' evaluations in text (Warriner et al., 2013). The words contained in the WRW range from those that may signal an evaluative reaction such as "exciting" or "worthwhile," but also words less likely to signal an evaluative reaction such as "aardvark" or "biography." Indeed, another major aim of the WRW is to provide ratings of a large

set of words to, for example, understand how different facets of those words affect how easily they are recognized (Kuperman, Estes, Brysbaert, & Warriner, 2014). As such, the WRW is focused relatively more on comprehensive word coverage, regardless of the evaluative nature of those words.

The WRW is similar to the EL in its approach to quantifying text as it also elicits normative ratings from participants for each word. In particular, the WRW consists of ratings for each word's valence, arousal, and dominance (Bradley & Lang, 1999; Osgood, Suci, & Tannenbaum, 1957). To construct the wordlist, researchers asked participants to judge each word on the basis of how they felt while reading it, each on a 1 to 9 scale, where the anchors respectively indicated *happy* to *unhappy* (valence), *excited* to *calm* (arousal), and *controlled* to *in control* (dominance).

The WRW measures an aspect of emotionality via its arousal dimension—a dimension that has been put forth as a key facet of emotion (e.g., Russell, 2003; Russell & Feldman Barrett, 1999). As evidence of the value of the WRW, past work has shown, for instance, that both its valence and arousal dimensions predict how quickly individuals are to recognize words that vary on these dimensions (Kuperman et al., 2014). Moreover, the WRW has also been used to provide insight into the nature of language and the biases it may contain (Warriner & Kuperman, 2015; see also Koch, Alves, Krüger, & Unkelbach, 2016).

Though arousal is a fundamental aspect of emotion, research indicates that emotion can be both high and low in arousal. For example, a person may feel excited or angry—both high arousal emotions—but can also feel pleased or depressed—low arousal emotions. Put simply, a word may be low in arousal but nonetheless indicative of emotionality (e.g., "pleased" or "sad"). Thus, the aim of the EL differs from the WRW as the EL seeks to measure emotionality in general. Indeed, we demonstrate that the emotionality of the word as assessed by the EL need not be the same as the level of arousal it is judged to elicit and that emotionality is uniquely associated with our outcomes of interest.

LIWC LIWC contains a number of ready-made wordlists, or dictionaries, that seek to measure different constructs of relevance to researchers. As a testament to its versatility, LIWC provides researchers with approximately 90 variables for each set of text it analyzes (Pennebaker et al., 2015). Researchers have used LIWC to provide evidence, for example, that greater linguistic similarity between individuals—in terms of, for instance, personal pronoun use—predicts increased mutual romantic interest in speed dating couples and, in a separate study, the relationship stability of couples three months later (Ireland et al., 2011; see Tausczik & Pennebaker, 2010, for a review).

LIWC has also been used to measure individuals' emotional reactions via what researchers have termed the *affect dictionary*, a category that is further divided into the *positive* and *negative emotion* subdictionaries (Pennebaker et al., 2015). For example, the positive emotion dictionary contains the words "amazing" and "okay" and the negative emotion dictionary contains the words "awful" and "inferior." If the target text contains the word "amazing" or "okay," then LIWC tallies each as a separate instance of positive emotion in the text. LIWC then calculates the percentage of positive (negative) words out of the total number of words used. Thus, LIWC's approach differs from both the EL and WRW in that it uses a word count approach to calculate valence (i.e., how often a set of words appears in the text), rather than through imputing normative values. A key difference of this approach is that whereas the EL and WRW measure differences amongst single words, LIWC counts all words as indicative of the same degree of emotionality. For example, LIWC's affect dictionary would treat a reviewer who "loved" a movie as identical to a reviewer who thought the movie was "worthwhile" to see—both would simply be treated as positive. In contrast, the EL treats "loved" as implying significantly greater emotionality. For an example directly comparing the EL and LIWC and their measurement of evaluative reactions, see Table 3.

Of interest, the positive and negative emotion dictionaries of LIWC have yet to be directly validated as measures of emotionality per se. For example, previous work has shown that participants used a greater percentage of words from LIWC's positive emotion dictionary when they were asked to write about a time when they felt amused, and a greater percentage of words from the negative emotion dictionary when they were asked to write about a time when they felt sad (Kahn, Tobin, Massey, & Anderson, 2007). However, no conditions were present in which participants were simply asked to write about positive or negative experiences that were unemotional. As such, it is possible that LIWC's affect dictionary may be more related to valence as opposed to emotionality. Indeed, although the affect dictionary of LIWC does contain words that are intuitively more emotional, such as "amazing" and "awful," it also contains words that are rather unemotional, such as "worthwhile" and "inferior." Taken together, it is ambiguous whether LIWC is related to emotion, per se, as opposed to the frequency of positive or negative words in the text. We examined this possibility in our analyses.[9]

[9] The newest version of LIWC also includes an *Emotional Tone* score, which is calculated as the difference between the positive emotion and negative emotion dictionaries that is then standardized (Cohn, Mehl, & Pennebaker, 2004; Pennebaker Conglomerates, Inc., n.d.). We utilized the positive and negative emotion dictionaries themselves in the present work given their longer history of use.

Table 3 Comparison of the Evaluative Lexicon 2.0 (EL 2.0) and the affect dictionary of LIWC

Text Example	Emotionality (EL 2.0)	Extremity (EL 2.0)	Affect (LIWC)
It was a nice movie. It was worthwhile to see.	4.03	3.12	20.00
I have a lot of fondness for this fun movie.	5.75	3.12	20.00
I loved this movie and the acting. It was amazing.	7.43	3.54	20.00

The underlined words are those present in both the EL 2.0 and LIWC wordlists. Numbers for the EL 2.0 are the average implied emotionality (out of 9.00) and extremity (out of 4.50) of the two words in each example. As calculated by LIWC, the numbers for LIWC are the percentage of words from its affect dictionary (two words) out of the total number of words in each example (ten total words).

Associations at the level of the words

As a first step in comparing the EL 2.0 to the other wordlists, we assessed the extent to which the different dimensions of each wordlist correlated with one another for those words that are common to each list. These analyses provide an initial indication of the extent to which the constructs measured by the different wordlists are similar or different. Given that the WRW seeks to cover as many words as possible, it shares a number of words with the EL 2.0 and LIWC. The WRW shares 894 words with the EL 2.0 and 371 with LIWC. The EL 2.0 and LIWC share 234 common words.[10]

Although the EL 2.0 and WRW provide disparate ratings of emotionality and arousal, respectively, we were able to provide a correlation between these two wordlists because they both utilize normative ratings for each word. Given that LIWC counts each word as a single instance of either positivity or negativity, however, we coded whether each word from LIWC was present in the positive (coded as "1") or negative (coded as "–1") emotion dictionaries. We report the correlations between the different wordlists and their ratings in Table 4, and provide a brief description of the primary findings here.

First, strong associations exist between the wordlists regarding the positivity of the words (rs ≥ .94). This result indicates that, of those words present on more than one list, strong agreement exists for whether the words are generally positive or negative.[11]

Second, in line with the initial work using the WRW (Warriner et al., 2013), the valence dimensions of the different wordlists were highly correlated with the dominance ratings of the WRW. As the WRW authors detailed when they introduced the wordlist, it is ambiguous whether dominance, as currently measured, is separable from valence (see also Stadthagen-Gonzalez et al., 2017). Our present work therefore focuses on the valence and arousal dimensions offered by the WRW for subsequent analyses.

Finally, although the emotionality, extremity, and arousal implied by the overlapping words are related, they appear to represent distinct constructs. The rather moderate association between emotionality and arousal in particular is in line with prior research that suggests that emotionality can be high or low in arousal (e.g., Russell, 2003; Russell & Feldman Barrett, 1999). Of note, the WRW arousal ratings are purported to reflect how calm to excited a participant felt when reading the word, whereas the EL emotionality ratings reflect the extent to which individuals are emotional in their evaluation. Thus, some caution should be taken in interpreting the correlation between these facets. Nevertheless, given their moderate correlation, these variables appear to measure related but separable constructs.

Taken together, these associations indicate that for those words that are common to the different wordlists, there is high agreement on the positivity or negativity implied by the words. Most important for our purposes, however, is that emotionality measured via the EL 2.0 appears to be distinct from the constructs measured via the other wordlists.

Differentiation of the wordlists on the measurement of emotionality

To further differentiate the constructs put forth by each tool, we examined the WRW and LIWC with regard to their assessment of emotionality using the 5.7 million product reviews we had obtained previously.

WRW If the arousal measured by the WRW is the same construct as emotionality, we should observe similar associations between this arousal measure and the emotion-signaling versus cognition-signaling phrases we identified previously. To that end, we calculated the proportion of times each of the words from the WRW was used with the emotion-signaling phrases compared to the cognition-signaling phrases [times used with emotion-signaling phrases/(times used with emotion-signaling phrases + times used with cognition-

[10] Note that these figures do not fully represent the number of words present in the LIWC wordlist due to its use of word stems. For instance, LIWC contains the word stem "livel" in the positive emotion dictionary. When LIWC encounters any word that begins with "livel," (e.g., "lively," "livelihood"), it counts that as an instance of positivity. Given that each of these words may imply different levels of valence, emotionality, or arousal from the EL and the WRW, they would not be able to be matched.

[11] For those readers who are interested, when calculating the extremity of the WRW on the basis of its valence scale, there was a strong correlation between the extremity of the words as rated in the WRW and the EL 2.0 [$r(892) = .67$, $p < .001$].

Table 4 Correlations between the dimensions of each wordlist at the level of the word

	Valence (EL 2.0)	Extremity (EL 2.0)	Emotionality (EL 2.0)	Valence (WRW)	Arousal (WRW)	Dominance (WRW)	Positive Emotion (vs. Negative) (LIWC)
Valence (EL 2.0)	–						
Extremity (EL 2.0)	– .10***	–					
Emotionality (EL 2.0)	– .13***	.47***	–				
Valence (WRW)	.96***	– .12***	– .10**	–			
Arousal (WRW)	– .02	.28***	.43***	– .19***	–		
Dominance (WRW)	.88***	– .08*	– .10**	.72***	– .18***	–	
Positive emotion (vs. negative) (LIWC)	.98***	.13*	– .11	.94***	– .06	.87***	–

* $p < .05$; ** $p < .01$; *** $p < .001$. The number of observations per cell, and thus the statistical significance, differs on the basis of the wordlists being compared: EL and WRW (894 common words), EL and LIWC (234), LIWC and WRW (371).

signaling phrases)]. As with the EL 2.0 analyses, those words that did not occur at all with these phrases (n = 1,211) and those words used fewer than ten times overall (n = 4,052) were not included. Moreover, we did not include the words "feel," "emotion," "emotional," "consider," or "believe" from the WRW given that they were the very words we were using to validate it. This left us with 8,647 words.

We first correlated the emotionality proportion (M = .62, SD = .13) with the arousal ratings of each word from the WRW, controlling for the extremity of the word. Though the correlation was significant and in the hypothesized direction, it was somewhat small in size [$r(8644) = .09$, $p < .001$]. Next, we directly compared the relation of the EL 2.0 and WRW to the emotionality proportion using the 725 words that were covered by both wordlists and that were used at least ten times. Specifically, we used a regression equation to predict the emotionality proportion from the extremity ratings from each wordlist as well as the implied emotionality from the EL 2.0 and the arousal ratings from the WRW. We found only an effect of EL 2.0 emotionality: the more emotional the word, the more it was used with emotion-signaling phrases [B = 0.04, $t(720) = 7.43$, $p < .001$]; conversely, this same term indicates that the more cognitive the word, the more it was used with cognition-signaling phrases. The arousal term from the WRW was non-significant [B = 0.001, $t(720) = .11$, p = .91] as were the extremity ratings from the EL 2.0 [B = – 0.001, $t(720) = .23$, p = .82] and WRW [B = – 0.009, $t(720)$ = 1.49, p = .14]. Taken together, these results indicate the EL 2.0 emotionality ratings are relatively closer in their relation to words we would expect to co-occur with more emotional reactions, and thus represent a construct that is separable from arousal as measured by the WRW.

LIWC As LIWC treats all words in its affect dictionary as indicative of the same level of emotionality, we were unable apply the same approach we had used for the WRW to LIWC. However, on the basis of both the past (Rocklage & Fazio, 2015, 2016) and present work, we have demonstrated the EL 2.0 as a valid measure of emotionality. Thus, we could examine the extent to which LIWC's affect measure was related to emotionality as measured via the EL 2.0 within the 5.7 million reviews we used. Moreover, as we discussed in the introduction to this section, it is important to consider whether the affect dictionary in LIWC is more related to emotionality per se or to the frequency of evaluative words present in the text. To assess this possibility, we quantified the extent to which each review contained evaluative language by using the frequency of EL 2.0 words used [number of EL 2.0 words in review/total words in review].

Across the 5 million reviews covered by both wordlists, controlling for the extremity of the reviews as measured by the EL 2.0, little association was observed between emotionality as measured via the EL 2.0 and the affect estimated by LIWC [$r(5005038) = .05$, $p < .001$]. Instead, LIWC's affect measure showed a stronger association with the frequency of evaluative words in a review [$r(5005039) = .78$, $p < .001$]. Moreover, the frequency of evaluative words itself does not appear to represent a proxy for emotionality; controlling for extremity, evaluative word frequency as measured by the EL 2.0 showed little association with EL 2.0 emotionality [$r(5005038) = .003$, $p < .001$].

Discussion This pattern of results suggests that the arousal dimension of the WRW, the affect dictionary of LIWC, and the emotionality dimension of the EL 2.0 measure different constructs. For LIWC in particular, its affect dictionary appears to be more closely related to the frequency of evaluative words present in a piece of text, and not necessarily to emotionality. Although it was possible that individuals who were more emotional would use a greater frequency of evaluative words, this was not the case in the current context.

Comparison of the wordlists in the prediction of self-reported evaluations

As a final comparison between the wordlists, we examined the ability of each wordlist to predict individuals' self-reported evaluations as expressed via their final star ratings. We first examined the WRW and LIWC wordlists by themselves. We then placed them into a single regression model with the EL 2.0, to examine the extent to which the different facets of each wordlist predicted individuals' evaluations. Finally, given that we were interested in the unique contributions of each wordlist, we used hierarchical regression to assess the change in variance accounted for (R^2) over and above the EL 2.0. We utilized the 5.7 million product reviews that contained words covered by all three wordlists, and were left with just over 5 million reviews.

WRW To quantify the reviews using the WRW, we followed the same approach as for the EL 2.0. Specifically, any word that fell above the midpoint of the valence scale of the WRW (5.00) was categorized as positive and any word that fell below the midpoint was categorized as negative. Any deviation from this midpoint signaled either greater positive or negative extremity. We then calculated weighted averages for positive extremity and negative extremity separately. Finally, we calculated a difference score as positive minus negative extremity. Arousal was calculated using the same approach (arousal to positivity minus arousal to negativity).

As one might expect, both greater positive (negative) extremity [$B = 0.35$, $t(5003692) = 380.90$, $p < .001$] and positive (negative) arousal [$B = 0.54$, $t(5003692) = 581.50$, $p < .001$] were related to more positive (negative) star ratings. These WRW variables accounted for 6.7% of the variance (i.e., $R^2 = .067$) in individuals' final ratings.

LIWC In a separate regression model, we entered the positive and negative emotion variables from LIWC as is. Both greater positive emotion [$B = 0.32$, $t(5003692) = 613.20$, $p < .001$] and greater negative emotion [$B = -0.33$, $t(5003692) = 620.30$, $p < .001$] predicted individuals' final ratings, as expected. These LIWC variables accounted for 16% of the variance (i.e., $R^2 = .16$) in the ratings.

Combined assessment of the wordlists We then entered the WRW, LIWC, and EL 2.0 variables into a single regression model. For the EL 2.0, we utilized the extremity, emotionality, and valence focus difference variables we detailed previously.

First, for the EL 2.0, we replicated the effects reported previously for extremity [$B = 0.49$, $t(5003687) = 438.24$, $p < .001$], valence focus [$B = 0.48$, $t(5003687) = 542.55$, $p < .001$], and emotionality [$B = 0.01$, $t(5003687) = 4.48$, $p < .001$].

Second, although WRW extremity and arousal predicted individuals' final ratings in the hypothesized manner when entered by themselves, these variables now predicted the opposite of what we might expect as more positive (negative) extremity [$B = -0.01$, $t(5003687) = 16.48$, $p < .001$] and arousal [$B = -0.01$, $t(5003687) = 9.54$, $p < .001$] were related to more negative (positive) overall ratings. Thus, in a regression model with the EL 2.0 and LIWC, the WRW appears to be less related to individuals' final judgments.[12]

Third, LIWC positivity [$B = 0.11$, $t(5003687) = 219.97$, $p < .001$] and negativity [$B = -0.07$, $t(5003687) = 142.18$, $p < .001$] were significant in the expected directions.

As we mentioned previously, individuals' most emotional reaction may be more indicative of the extent of their emotionality toward the product. To this end, we estimated this same regression model using individuals' most emotional positive reaction minus their most emotional negative reaction, as measured via the EL 2.0. As before, we found much stronger effects of emotionality [$B = 0.11$, $t(5003687) = 105.31$, $p < .001$], over and above the effects of the other EL 2.0 dimensions as well as of the other wordlists.

In terms of variance accounted for, the three EL 2.0 variables accounted for 34.9% of the variance in individuals' final star ratings—over five times the amount accounted for by WRW and over twice that of LIWC when each was entered alone. Moreover, when entered sequentially using hierarchical regression, above and beyond the EL 2.0, the WRW (R^2 change over and above the EL 2.0 = .001) and LIWC (R^2 change over and above both the EL 2.0 and WRW = .008) variables accounted for less than 1% more unique variance.

Discussion These results demonstrate that when used separately in a predictive model, the WRW and LIWC predict individuals' final ratings as we might expect. However, when entered into a model alongside the EL 2.0, the EL 2.0 provided the best predictive ability. These results make sense in light of the aims of each of these tools. Whereas the WRW and LIWC each seek to provide utility apart from understanding individuals' evaluations, the EL was specifically designed to measure evaluations.

General discussion

Using an iterative, data-driven approach, trained judges, and a large corpus of 9 million online reviews, we constructed and

[12] To ensure these results for the WRW were not specific to the approach we used, we also calculated the average valence as well as the average arousal implied by each review. We then used a valence by arousal interaction to predict the final star ratings. With the same regression model used above, we found similar results such that greater arousal to positivity predicted more negative star ratings [$B = -0.01$, $t(5003686) = 23.30$, $p < .001$] and greater arousal to negativity predicted more positive star ratings [$B = 0.05$, $t(5003686) = 80.12$, $p < .001$]. Thus, although these particular relationships were not expected, they were consistent across different approaches.

introduced the Evaluative Lexicon 2.0 (EL 2.0). To create the EL 2.0, we identified those words that were indicative of individuals' evaluative reactions and that were used relatively consistently across a wide range of topics. This led to an expansion of the EL 1.0's 94 adjectives to 1,541 words—an increase of nearly 1600%. We then validated the EL 2.0 across four separate sets of reviews, or 5.7 million new reviews in total. We demonstrated that phrases we anticipated to co-occur with more emotional reactions (e.g., "I feel") versus more cognitive reactions (e.g., "I believe") indeed accompanied the more emotional versus cognitive words. Moreover, we demonstrated that the EL 2.0 predicted individuals' final star ratings across these online reviews.

We also differentiated the EL 2.0 from other approaches assumed to capture an aspect of emotionality: the WRW and LIWC. Our results indicated that emotionality as measured via the EL 2.0 was related to phrases signaling emotional reactions whereas the arousal dimension of the WRW was relatively unrelated to these phrases. For LIWC, when correlating the EL 2.0 with the LIWC's affect dictionary across 5 million reviews, we found that LIWC showed comparatively little association with emotionality as measured via the EL 2.0. Instead, LIWC showed a stronger association with the frequency of evaluative words within the text, which suggests that LIWC may be relatively more sensitive to the frequency of evaluative words within text rather than to emotionality per se. Taken together, these results indicate that the EL 2.0 is relatively unique in its measurement of emotionality in text. Finally, we found that the EL 2.0 was also a better predictor of individuals' final summary ratings—a property consistent with its explicit aim to measure evaluations.

Quantification of natural text using the Evaluative Lexicon 2.0

In this section we offer researchers guidance on how they may best quantify text using the EL 2.0. For the analyses in the present article, we utilized a simple difference score (e.g., emotionality to positive valence minus emotionality to negative valence) to demonstrate the utility of the EL 2.0 in a straightforward manner. Although this general approach is useful, it is important for researchers to consider their goals in order to determine the quantitative approach suited to their needs.

To illustrate, Rocklage and Fazio (2015, Study 3) were interested in the effects of emotionality when reviewers indicated they had both positive and negative reactions to their products (i.e., the reviewers expressed ambivalence) as well as when they had either just positive or just negative reactions (i.e., the reviewers expressed univalence). To that end, they analyzed each set of reviews separately using different approaches. For the ambivalent reviews, they used the difference scores approach reported here. However, for the univalent reviews, they created a dichotomous variable to indicate whether the review was positive or negative and then used two interactions to model the effects of extremity and emotionality: extremity by valence and emotionality by valence (see Rocklage & Fazio, 2015, Study 3, for further details). These approaches allowed them to make specific statements regarding each type of review.

Researchers should also consider additional factors beyond simply quantifying their text and predicting their variables of interest. As an example from the present work, online reviewers are expected to provide a rationale for their evaluative reactions, and thus may introduce additional details about the product that can be rather cognitive and unemotional. Thus, reviewers' most emotional reaction was a better predictor of their final summary judgments than was the average emotionality implied across their review. This example illustrates the importance of considering the context in which the text was written and how researchers may adapt the tool given their knowledge of the situation (see also Rocklage & Fazio, 2015, 2016).

Although we have focused on the power of the EL to measure emotionality, it also captures the implied valence and extremity of individuals' words. As we outlined in the introduction and have demonstrated in this work, emotionality and extremity are related but separable constructs. One can have a positive evaluation of an item and believe it is "helpful," but, on the other hand, provide an equally positive but more emotional evaluation of an item as "enjoyable" to use. These different dimensions can be used to provide a more complete picture of one's phenomenon and thus we urge researchers to consider the relevance of each dimension for their hypotheses. For instance, evidence exists that readers often find more extreme online reviews to be more helpful. However, the emotionality of the review can have the same or even opposite effect on readers, depending on the type of product the review is about. Specifically, readers judge emotional reactions to be more helpful for products such as movies, but they judge more unemotional, cognitive reactions as more helpful for products such as blenders (Rocklage & Fazio, 2017). In this case, for products like blenders, the opposing effects of extremity and emotionality could lead each term to be non-significant when analyzed by itself, but strongly significant in opposite directions when analyzed together.

One approach we recommend against is to use the valence and extremity variables as-is within the same statistical model. The calculation of extremity is based directly on the valence variable. Once the direction of the extremity is known as either positive or negative in valence, this variable largely becomes redundant with valence. To this end, one can take the present approach of creating difference variables of positive minus negative extremity and emotionality. As we mentioned previously, another approach, for those reviews that use just positive or just negative words, is to create a dichotomous variable to categorize each review as either positive or negative. Then, this dichotomous valence variable can be interacted with the

extremity variable and the emotionality variable to assess the effects of these variables and, for example, how one's effects may differ by valence (see Rocklage & Fazio, 2015, Study 3; Rocklage & Fazio, 2016).

Additional uses of the Evaluative Lexicon 2.0 in natural text

In the present work, we sought to construct, validate, and demonstrate the utility of the EL 2.0. As such, we focused on a subset of uses of the tool. However, researchers can use the EL 2.0 for a variety of purposes. As we have already noted, the EL 2.0 can identify whether individuals express mixed reactions (i.e., ambivalence) by the detection of both positive and negative words within the same text. In addition, the EL variables can be used to predict outcomes for individuals as done here, or even aggregated to assess macro-level outcomes. For example, future research could track how people communicate on a large scale via word-of-mouth communications (e.g., Berger & Milkman, 2012), predict the success of brands and products in the marketplace (e.g., Bagozzi, Gopinath, & Nyer, 1999; Carroll & Ahuvia, 2006; Cohen, Pham, & Andrade, 2008), or use nationwide variability in individuals' evaluative reactions to predict outcomes at the county or state level (e.g., Mitchell, Frank, Harris, Dodds, & Danforth, 2013).

In the present work, we utilized online reviews which offered the advantage of being accompanied by a final quantitative judgment in the form of individuals' star ratings. As such, we were able to use this rating to validate different aspects of the EL 2.0 as well as compare it with other existing tools. Moreover, these reviews are also advantageous as they allowed us to construct a relatively stable measure of evaluations across diverse topics. Of course, a limitation is that we did not examine the EL 2.0's ability to quantify text in other online contexts—such as Twitter—or offline contexts—such as political speeches. Indeed, we believe future research would benefit from a deeper exploration of the capability of the EL 2.0 to predict outcomes in additional contexts.

On this matter, an appealing quality of the EL is its potential to quantify individuals' reactions even in short pieces of text, such as tweets from Twitter or posts on Facebook. Though not investigated in the present work, preliminary research suggests that the EL 2.0 can quantify text in these or other online contexts as well e.g., Twitter; Rocklage, Rucker, & Nordgren, 2017). Moreover, as has been shown in other work, reactions on social media are predictive of real-world outcomes such as box office revenue for movies (Asur & Huberman, 2010) and the performance of the stock market (Bollen, Mao, & Zeng, 2011). Thus, these contexts represent promising avenues for future research.

Availability

A final advantage of the EL 2.0 is that it is free and easy to obtain. The EL 2.0 is available as an Excel file and is provided in the supplementary materials that accompany this article. It can also be obtained from www.evaluativelexicon.com. This file contains the normative emotionality, extremity, and valence implied by each word as rated by the large set of judges reported here.

Conclusion

Over the past decade, researchers have found themselves with the exciting prospect of studying human behavior on a truly massive scale. To help researchers take advantage of this opportunity, we have introduced the EL 2.0—a validated measure of the emotionality, extremity, and valence of individuals' evaluative reactions in natural text. As such, the EL 2.0 provides researchers with the opportunity to capture evaluative reactions both in the laboratory and "in the wild."

References

Abelson, R. P., Kinder, D. R., Peters, M. D., & Fiske, S. T. (1982). Affective and semantic components in political person perception. *Journal of Personality and Social Psychology*, *42*, 619–630. doi:https://doi.org/10.1037/0022-3514.42.4.619

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media. In 2010 IEEE/ACM International Conference on Web Intelligence-Intelligent Agent Technology (WI-IAT)* (Vol. 1, pp. 492–499). Los Alamitos, CA, USA: IEEE Computer Society. doi:https://doi.org/10.1109/WI-IAT.2010.63

Bagozzi, R. P., & Burnkrant, R. E. (1979). Attitude organization and the attitude–behavior relationship. *Journal of Personality and Social Psychology*, *37*, 913–929. doi:https://doi.org/10.1037/0022-3514.37.6.913

Bagozzi, R. P., Gopinath, M., & Nyer, P. U. (1999). The role of emotions in marketing. *Journal of the Academy of Marketing Science*, *27*, 184–206. doi:https://doi.org/10.1177/0092070399272005

Batra, R., & Ahtola, O. T. (1991). Measuring the hedonic and utilitarian sources of consumer attitudes. *Marketing Letters*, *2*, 159–170. doi:https://doi.org/10.1007/BF00436035

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*, 192–205. doi:https://doi.org/10.1509/jmr.10.0353

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media, Inc.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*, 1–8. doi:https://doi.org/10.1016/j.jocs.2010.12.007

Bradley, M. M., & Lang, P. J. (1999). *Affective Norms for English Words (ANEW): Instruction manual and affective ratings* (Technical Report No. C-1). Gainesville, FL: University of Florida, NIMH Center for Research in Psychophysiology.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the

introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990. doi:https://doi.org/10.3758/BRM.41.4.977

Carroll, B. A., & Ahuvia, A. C. (2006). Some antecedents and outcomes of brand love. *Marketing Letters*, *17*, 79–89. doi:https://doi.org/10.1007/s11002-006-4219-2

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*, 345–354. doi:https://doi.org/10.1509/jmkr.43.3.345

Cicero, M. T. (1986). *On oratory and orators* (J. S. Watson, Trans.). Carbondale, IL: Southern Illinois University Press.

Cohen, J. B., Pham, M. T., & Andrade, E. B. (2008). The nature and role of affect in consumer behavior. In C. P. Haugtvedt, P. M. Herr, & F. R. Kardes (Eds.), *Handbook of consumer psychology* (pp. 297–348). Mahwah, NJ: Erlbaum.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, *15*, 687–693. doi:https://doi.org/10.1111/j.0956-7976.2004.00741.x

Crites, S. L., Fabrigar, L. R., & Petty, R. E. (1994). Measuring the affective and cognitive properties of attitudes: Conceptual and methodological issues. *Personality and Social Psychology Bulletin*, *20*, 619–634. doi:https://doi.org/10.1177/0146167294206001

Eagly, A. H., Mladinic, A., & Otto, S. (1994). Cognitive and affective bases of attitudes toward social groups and social policies. *Journal of Experimental Social Psychology*, *30*, 113–137. doi:https://doi.org/10.1006/jesp.1994.1006

Fabrigar, L. R., & Petty, R. E. (1999). The role of the affective and cognitive bases of attitudes in susceptibility to affectively and cognitively based persuasion. *Personality and Social Psychology Bulletin*, *25*, 363–381.

Haddock, G., Zanna, M. P., & Esses, V. M. (1993). Assessing the structure of prejudicial attitudes: The case of attitudes toward homosexuals. *Journal of Personality and Social Psychology*, *65*, 1105–1118. doi:https://doi.org/10.1037/0022-3514.65.6.1105

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834. doi:https://doi.org/10.1037/0033-295X.108.4.814

Hirschman, E. C., & Holbrook, M. B. (1982). Hedonic consumption: Emerging concepts, methods and propositions. *Journal of Marketing*, *46*, 92–101. doi:https://doi.org/10.2307/1251707

Huskinson, T. L. H., & Haddock, G. (2004). Individual differences in attitude structure: Variance in the chronic reliance on affective and cognitive information. *Journal of Experimental Social Psychology*, *40*, 82–90. doi:https://doi.org/10.1016/S0022-103100060-X

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, *22*, 39–44. doi:https://doi.org/10.1177/0956797610392928

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In M. Najork, A. Broder, & S. Chakrabarti (Eds.), *Proceedings of the 2008 International Conference on Web Search and Web Data Mining* (pp. 219–230). New York, NY: ACM Press. doi:https://doi.org/10.1145/1341531.1341560

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the linguistic inquiry and word count. *American Journal of Psychology*, *120*, 263–286. doi:https://doi.org/10.2307/20445398

Koch, A., Alves, H., Krüger, T., & Unkelbach, C. (2016). A general valence asymmetry in similarity: Good is more alike than bad. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1171–1192. doi:https://doi.org/10.1037/xlm0000243

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, *143*, 1065–1081. doi:https://doi.org/10.1037/a0035669

Lavine, H., Thomsen, C. J., Zanna, M. P., & Borgida, E. (1998). On the primacy of affect in the determination of attitudes and behavior: The moderating role of affective-cognitive ambivalence. *Journal of Experimental Social Psychology*, *34*, 398–421.

Lazarus, R. S. (1982). Thoughts on the relations between emotion and cognition. *American Psychologist*, *37*, 1019–1024. doi:https://doi.org/10.1037/0003-066X.37.9.1019

Luca, M. (2011). *Reviews, reputation, and revenue: The case of Yelp.com* (Working Paper No. 12-016). Harvard Business School, Cambridge, MA. Available at www.hbs.edu/faculty/Pages/item.aspx?num=41233

McAuley, J., Pandey, R., & Leskovec, J. (2015). Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. doi:https://doi.org/10.1145/2783258.2783381

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, *106*, 3–19. doi:https://doi.org/10.1037/0033-295X.106.1.3

Miller, G. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*, 39–41. doi:https://doi.org/10.1145/219717.219748

Mischel, W., & Shoda, Y. (1995). A cognitive–affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*, 246–268. doi:https://doi.org/10.1037/0033-295X.102.2.246

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLOS ONE*, *8*, e64417. doi:https://doi.org/10.1371/journal.pone.0064417

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Pennebaker Conglomerates, Inc. (n.d.). Interpreting LIWC output (Web page). Retrieved June 21, 2017, from http://liwc.wpengine.com/interpreting-liwc-output/

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Retrieved from https://repositories.lib.utexas.edu/handle/2152/31333

Petty, R. E., Fabrigar, L. R., & Wegener, D. T. (2003). Emotional factors in attitudes and persuasion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Oxford handbook of the affective sciences* (pp. 752–772). New York, NY, US: Oxford University Press.

Pham, M. T. (2007). Emotion and rationality: A critical review and interpretation of empirical evidence. *Review of General Psychology*, *11*, 155–178. doi:https://doi.org/10.1037/1089-2680.11.2.155

Rocklage, M. D., & Fazio, R. H. (2015). The Evaluative Lexicon: Adjective use as a means of assessing and distinguishing attitude valence, extremity, and emotionality. *Journal of Experimental Social Psychology*, *56*, 214–227. doi:https://doi.org/10.1016/j.jesp.2014.10.005

Rocklage, M. D., & Fazio, R. H. (2016). On the dominance of attitude emotionality. *Personality and Social Psychology Bulletin*, *42*, 259–270. doi:https://doi.org/10.1177/0146167215623273

Rocklage, M. D., & Fazio, R. H. (2017). *The phenomenal disjunction: Emotionality for ourselves versus others.* Manuscript submitted for publication.

Rocklage, M. D., Rucker, D. D., & Nordgren, L. F. (2017). *The Evaluative Lexicon: Capturing attitude valence, extremity, and emotionality "in the wild."* Paper presented at the Conference of the Society for Personality and Social Psychology, San Antonio, Texas.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*, 145–172. doi:https://doi.org/10.1037/0033-295X.110.1.145

Russell, J. A., & Feldman Barrett, L. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the

elephant. *Journal of Personality and Social Psychology*, *76*, 805–819. doi:https://doi.org/10.1037/0022-3514.76.5.805

Shiv, B., & Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research*, *26*, 278–292. doi:https://doi.org/10.1086/jcr.1999.26.issue-3

Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, *49*, 111–123. doi:https://doi.org/10.3758/s13428-015-0700-2

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*, 24–54. doi:https://doi.org/10.1177/0261927X09351676

van den Berg, H., Manstead, A. S. R., van der Pligt, J., & Wigboldus, D. H. J. (2006). The impact of affective and cognitive focus on attitude formation. *Journal of Experimental Social Psychology*, *42*, 373–379. doi:https://doi.org/10.1016/j.jesp.2005.04.009

Voss, K. E., Spangenberg, E. R., & Grohmann, B. (2003). Measuring the hedonic and utilitarian dimensions of consumer attitude. *Journal of Marketing Research*, *40*, 310–320. doi:https://doi.org/10.1509/jmkr.40.3.310.19238

Wang, H., Lu, Y., & Zhai, C. (2011). Latent aspect rating analysis without aspect keyword supervision. In C. Apte, J. Ghosh, & Padhraic Smyth (Eds.), *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 618–626). New York, NY, USA: ACM Press. doi:https://doi.org/10.1145/2020408.2020505

Warriner, A. B., & Kuperman, V. (2015). Affective biases in English are bi-dimensional. *Cognition and Emotion*, *29*, 1147–1167. doi:https://doi.org/10.1080/02699931.2014.968098

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191–1207. doi:https://doi.org/10.3758/s13428-012-0314-x

Williams, P., & Drolet, A. (2005). Age-related differences in responses to emotional advertisements. *Journal of Consumer Research*, *32*, 343–354. doi:https://doi.org/10.1086/497545

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*, 151–175. doi:https://doi.org/10.1037/0003-066X.35.2.151