Published in final edited form as:
Commun Methods Meas. 2011 December ; 5(4): 275–296. doi:10.1080/19312458.2011.624489.

Automating Content Analysis of Open-Ended Responses: *Wordscores* and *Affective Intonation*

Young Min Baek, **Joseph N. Cappella**, and **Alyssa Bindman**
University of Pennsylvania, Annenberg School for Communication

Abstract

This study presents automated methods for predicting valence and quantifying valenced thoughts of a text. First, it examines whether *Wordscores*, developed by Laver, Benoit, and Garry (2003), can be adapted to reliably predict the valence of open-ended responses in a survey about bioethical issues in genetics research, and then tests a complementary and novel technique for coding the number of valenced thoughts in open-ended responses, termed *Affective Intonation*. Results show that *Wordscores* successfully predicts the valence of brief and grammatically imperfect open-ended responses, and *Affective Intonation* achieves comparable performance to human coders when estimating number of valenced thoughts. Both *Wordscores* and *Affective Intonation* have promise as reliable, effective, and efficient methods when researchers content-analyze large amounts of textual data systematically.

One of the core methodological issues in communication research is the reliable and efficient coding of textual data, whether texts are formally offered in published documents (e.g., newspaper articles) or less formal verbal texts (e.g., open-ended responses or transcripts of interactions). Automated content analytic methods have received a great deal of attention from computer scientists and social scientists in many disciplines (Hopkins & King, 2009; Laver, Benoit, & Garry, 2003; Pang & Lee, 2008). One recent achievement in the field is *Wordscores*,[1] a technique that was developed in political science to code the ideological tone of formal political texts, such as party manifestos (Laver et al., 2003). Unlike a number of methods developed in computational linguistics, *Wordscores* is relatively easy for social scientists to use and has satisfactory validity and reliability (Klemmensen, Hobolt, & Hansen, 2007; Lowe, 2008) because it does not require any distributional assumptions about the words used.[2]

Correspondence should be addressed to: Young Min Baek, Annenberg School of Communications, 3620 Walnut Street, Philadelphia, PA 19104. ybaek@asc.upenn.edu.

[1]While *Wordscores* with upper letter of *W* is used as a methodological approach, *wordscores* with lower letter of *w* is used to mention scaled value of words generated from reference texts.

[2]There are two statistical approaches: (1) the unsupervised approach and (2) the supervised approach (Duda, Hart, & Stork, 2000). *Wordscores* is based on the supervised approach. Unlike *Wordscores*, the unsupervised approach does not rely on prior human interpretations of selected textual data. Like cluster analysis, the unsupervised approach summarizes several groups of textual data, based on words' frequencies, positions, and co-occurrence pattern with other words. In general, computer scientists or computational linguistics are more interested in developing the unsupervised approach because it does not demand any external knowledge of textual data and works in any kind of textual domain. However, most social scientific research questions deal with predetermined topics with theoretical concerns or predictions, making the social scientific research a viable fit with the supervised approach.

Although *Wordscores* is known to be a reliable method for formal documents, it has not been applied to informal, nonpolitical texts generated by ordinary people. Therefore, we begin by testing whether *Wordscores* successfully predicts the valence of short, informal, nonpolitical texts generated by representative groups of respondents. We also investigate the benefits of expanding *Wordscores* to *Affective Intonation*, our modified scaling application of *Wordscores*, so that it can be used as a tool to predict the number of *valenced thoughts or reasons* contained in a text. These variables are important outcomes in many contexts in communication research and the social sciences more generally. In persuasion and attitude change, thought-listing (Brock, 1967; Petty & Cacioppo, 1986; Petty & Wegener, 1998) has figured as both a significant outcome and mediating variable (Petty, Brinol, & Tormala, 2002). Successful persuasive communications generate more positive than negative thoughts; those with more negative thoughts are linked to unsuccessful or even boomerang effects. A related measure employed in large scale political communication research goes by the name of "considerations"; respondents are asked to list what they considered, pro and con, in their stated opinion (Cappella, Price, & Nir, 2002). Considerations are in the same class of variables as thoughts but have been used less as a diagnostic tool for persuasive messages than as a basis for quality of expressed opinions. The open-ended item seeks to elicit positive and negative thoughts that respondents have about candidates and policies, particularly in electoral contexts (Price, Nir, & Cappella, 2005).

Conceptual work by Price and Neijens (1997) has also led to the development of measures of quality of public opinion called argument repertoire (Cappella et al., 2002). These measures have successfully predicted participation in deliberative contexts and differentiate those who have participated in online deliberation from those who have not (Price & Cappella, 2002). Although coding procedures in and of themselves are not burdensome, they require training of human coders and careful assessment and reassessment of coders to assure adequate reliability of content over time. For very large samples of respondents or multiple tests, the coding burden can quickly become resource intensive and make researchers reluctant to employ these techniques despite their utility.

Other studies have sought to link temporal trends in news coverage to trends in social and behavioral outcomes. Such studies usually require lengthy time-series for the chosen outcome as well as content analysis of news articles. Research of this type has addressed political behavior (Shah, Watts, Domke, & Fan, 2002), risky decisions (Romantan, 2004), public policy outcomes (Yanovitsky & Bennett, 1999), drug use (Fan & Halloway, 1994), and risky health outcomes (Yanovitsky & Stryker, 2001). These kinds of studies require efficient content analytic procedures due to the magnitude of the data collected and categorized as a part of the time series. In some cases, researchers have been able to sidestep the assessment of valence of the content by assuming, for example, that news coverage of a particular topic will not be favorable (e.g., drug use). In those cases, the frequency of treatment of the topic alone is indicative of the tone. While such an assumption might make sense for certain topics such as "drug use" in mainstream news sources, as scholars seek out more controversial topics or broaden analyses to blogs or websites, texts are likely to return a much more diverse set of contents that include both positively and negatively valenced segments.

Consequently, efficient content analysis that can distinguish and enumerate valence of text is a core agenda item in communication research. If coding of a variety of texts can be solved computationally and manual labor can be significantly reduced, it will give researchers more latitude and incentive to investigate the extent to which the content of textual messages affects social, behavioral and psychological outcomes or to employ open-ended responses as indicators of underlying psychological states and cognitive models.

In the first part of the article, we present two automated content analytic methods: Laver et al.'s *Wordscores* (2003) for predicting valence of a text and *Affective Intonation*, our newly suggested scaling alternative to *Wordscores*, for predicting the number of valenced thoughts in the text. In the second part, we introduce textual data that are automatically content analyzed in order to illustrate the potential utility of *Wordscores* and *Affective Intonation*. In the third part, the *Wordscores* and *Affective Intonation* are tested to assess (1) *concurrent validity* and (2) compare *their predictive validity* against that of manual coding in a context where texts are short, informal, and generated without careful editing. In the last section, we address how the procedures can be modified to apply both *Wordscores* and *Affective Intonation* to a variety of standard content analyses in communication research.

WORDSCORES AND AFFECTIVE INTONATION

Wordscores: Predicting Valence of Target Texts by Scaling Words in Reference Texts[3]

Wordscores is different from other content-analytic methods (Laver et al., 2003), such as traditional human content analysis (Krippendorff, 2004) or an established dictionary-based computerized content analysis (Laver & Garry, 2000; Pennebaker, Francis, & Booth, 2001; Popping, 2000; Schrodt & Gerner, 1994). *Wordscores* treats texts "as collections of word data" (Laver et al., 2003, p. 312), rather than something to be "interpreted" by intelligent agents (Krippendorff, 2004).

Five stages are required: (1) identifying the affective dimension[4] (e.g., positive – negative in open-ended responses), (2) selecting the reference texts and their category (e.g., negative = −1, and positive = +1), (3) generating wordscores from reference texts (see below for details), (4) applying wordscores to target texts whose valence scores are unknown (or treated as unknown for testing purposes as they are here), and (5) estimating target texts by averaging wordscores in the text. The full mathematical details of *Wordscores* have been discussed elsewhere (Laver et al., 2003; Lowe, 2008; Monroe & Schrodt, 2008) and so will only be summarized here.

The first stage of *Wordscores* is to set a clear goal for coding. For example, in the case of public opinion research, scholars may focus on attitudinal valence of an issue. In our data, responses can be located on the continuum of "positivity-to-negativity" towards an issue pertinent to voluntary participation in genetics research.

[3]The original developers (Laver et al., 2003) called target texts "virgin texts." However, we deliberatively avoid the term to avoid connotations related to sexuality and gender.

[4]The original developers (Laver et al., 2003) may favor "unidimensional policy position of a political document."

The second step selects reference texts as the basis for scaling the valence of words. The reference text sample is important (Laver et al., 2003) because *Wordscores* assume that words appearing in the reference texts and the target texts are similar (Lowe, 2008). This correspondence determines the success of automated coding performance (Klemmensen et al., 2007). The developers of *Wordscores* selected subsets of sampled texts as reference texts and let experts categorize these subsets. Instead of experts' judgments, researchers can employ evaluations by any reliable source, including trained coders, aggregate scores from naive evaluators, or even "self-evaluations" as is sometimes done in thought-listing where people categorize their own thoughts as positive or negative. Any technique that results in clear and unequivocal valence of segments of text can be used for the *Wordscores'* algorithms. In the test carried out in this article, we randomly select a small portion of people's positive (i.e., supportive) (+1.00) and negative (i.e., opposed) thoughts (−1.00) and treat them as reference texts while the remaining responses are treated as target texts. To test the predictive value of the proportion of reference texts needed, we selected a range of samples from 1% (with the remaining 99% treated as target texts) to 50% (50% treated as target texts) of the total responses.

The third step estimates the score of a word, termed wordscore, based on its relative frequency in positive and/or negative reference texts where the valence of segments is known in advance. The result is a list of words that primarily function as "positive" words and another list that serve as "negative" words in their frequencies within *a priori* valenced texts. These words can be thought of as diagnostic sign posts or signals that will later serve as the basis for estimating valence of target text. Of course, some words will function as equally frequent in negative and positive texts and so will not be diagnostic for textual valence.

Notice there is no syntactic structure employed as the procedure essentially treats a text as a "bag of words" functioning primarily as indicators of positive valence or negative valence. Although this approach does injustice to the nature of linguistic communication, it is simple and efficient. If it is also effective for the purposes of coding, then the "bag of words" assumption is worth making.

Specifically, the wordscore is defined as follows[5] (Laver et al., 2003, pp. 315–316):

$$\begin{aligned}\text{Probability}_{\text{wr}} &= \frac{\text{Frequency}_{\text{wr}}}{\sum_r \text{Frequency}_{\text{wr}}} \\ \text{wordscore}_{\text{w}} &= \sum_{\text{r}} \text{Probability}_{\text{wr}} \times \text{Value}_{\text{r}}\end{aligned} \quad (1)$$

where *w* denotes specific word in reference text; *r* denotes reference text previously designated as positive or negative; $Value_r$ denotes the assigned value of the reference text (e.g., +1 = positive text, and −1 negative text). For example, assume that the word "religion" appeared 10 times in positive reference text and 190 times in a negative one. Then the probability of religion in positive reference text (i.e., $Probability_{w=religion\&r=positive}$) will be $.05\ (=\frac{10}{10+190})$, and "the probability of religion in negative reference text (i.e.,

[5]Laver et al. (2003) used more general form of formula, and thus our notation is slightly different from that of the developers of *Wordscores*.

Probability$_{w=religion\&r=negative}$)" will be .95 ($=\frac{190}{10+190}$). Using these probabilities, we can calculate the wordscore of "religion" (=.90) as follows:

$$\text{wordscore}_{\text{w=religion}}=\text{Probability}_{\text{w=religion \& r=negative}}\times\text{Value}_{\text{r=negative}}+\text{Probability}_{\text{w religion \& r=positive}}\times\text{Value}_{\text{r=positive}}$$
$$=(.95)\times(-1.00)+(.05)\times(1.00)=-.90$$

The fourth step is the application of a list of wordscores to words in the target texts. For example, if a target text contains the word "religion," then *Wordscores* assumes that "religion" contributes "−.90" to the target text's predicted valence. If a target text contains a word that was not observed in any of reference texts, then the word is treated as "missing" and is not included when predicting its valence.

As the last step, a target text's valence is estimated by averaging all valid wordscores in the text. For example, assume two responses (one is positive, and the other is negative) with wordscores in parentheses for each word:

Positive response:	I(.01) think(.02) genetic(.04) tests (−.01) will(−.01) improve(.77) medical(.83) treatment(.73) and(.05) may(−.03) help(.89) sick(.96) people(.40).
Negative response:	I(.01) think(.02) genetic(.04) tests (−.01) will(−.01) endanger(−.73) privacy(−.61) information(.10) and(.05) may(−.03) discriminate(−.84) racial(−.81) minorities(−.80).

In the above cases, the positive response will be .33, while the negative response will be −.29. Finally, *Wordscores* locates target texts using a "rescaling" technique to adjust the "variance shrinkage" (Lowe, 2008, p. 359) because wordscores of nonsubstantive words (e.g., 'a/an' or 'the') are close to zero, meaning the predicted valence of target texts tends unnecessarily close to the zero. There are two types of rescaling transformations—one is suggested by the developers (Laver et al., 2003) and the other by Martin and Vanberg (2008). Here we apply the original developers' rescaling technique.[6] After predicting valence of responses in target texts, we treated any response whose value is positive as a "positive" response and those with negative value as a "negative" response (i.e., no threshold is employed). When a target text comprises words not observed in reference texts (i.e., no valid wordscores in the target text), we treat those responses as "unknown," that is, the equivalent of missing, because the text cannot be predicted from words that do not appear in reference texts. So the predicted valence of responses has three categories (i.e.,

[6]The Laver et al.'s transformation procedure (2003) is

$$P_t^*=(P_t-\overline{P}_T)\left(\frac{SD_R}{SD_T}\right)+\overline{P}_T$$

where P_t is raw score for a text *t;* $\overline{P_T}$ is the averaged wordscores in a set of target texts *T*, and SD_R and SD_T are the standard deviations of the reference and target text scores, respectively.

As shown in the transformation procedure, the transformed score for a text *t* (i.e., P_t^*) readjusts the raw score P_t by calculating a text *t*'s relative distance from the mean score of target texts after adjusting standard deviations between reference texts and target texts. Here we treat P_T as "zero" because it is theoretically simple, and also our division between reference texts and target tests are random, implying that the expected value of P_T is 0.

positive, negative, and unknown), although the real valence of responses is either positive or negative (two categories).

Affective Intonation: Predicting the Valenced Thoughts in Target Texts Using *Wordscores*

Wordscores is a good automated content analytic method for predicting the valence of a text, especially when a large amount of coding is necessary. However, *Wordscores* is not appropriate to distinguish "more thoughtful responses" from "less thoughtful responses" because it *averages* observed wordscores in a target text. Therefore, conventional *Wordscores* is useful in predicting valence but not the number of valenced thoughts (as stated reasons). When researchers are interested in questions not just about the direction of texts but also about the degree of cognitive elaboration in those texts, then the number of written responses that function as reasons is the focus (Petty & Cacioppo, 1986; Cappella et al., 2002).

Consider the positive and negative texts employed earlier as follows: "I think genetic tests will improve medical treatment and may help sick people" and "I think genetic tests will endanger privacy information and may discriminate against racial minorities." Ordinary manual coding of thoughts or reasons in thought-listing or argument repertoire would detect two positive reasons in the first response (i.e., {improve, medical, treatment}, {help, sick}) and two negative reasons in negative response ({endanger, privacy}, {discriminate, racial, minorities}) because human coders understand that those words mainly represent positively or negatively valenced reasons in an issue discourse. Figure 1 graphs the wordscores of each word in the two texts seriatim. Two hills occur in the positive response (solid line) and two valleys in the negative (dotted line) paralleling the positive and negative reasons. This oscillation of wordscores in a response is what we are calling *Affective Intonation*, akin to phonetic intonation in speech.

The *Affective Intonation* procedure aims to mimic the human coders' judgments by ordering each word's wordscore on its sequential location in an open-ended response and detecting changes in these patterns. *Affective Intonation* starts with a word's wordscore as defined in Equation 1. Words with a stronger positive (or negative) wordscore can be considered a weightier indicator of a text's valenced reason. For example, it is plausible that a word whose wordscore is .90 is a stronger indicator of positive text than a word whose wordscore is .10. Also, wordscores around zero are weaker indicators because they appear in both positive and negative reference texts with similar probability. Our basic claim is that *Affective Intonation* operationalizes wordscores as measures of valenced reasons in a target text, meaning that *Affective Intonation* is a modified scaling alternative to *Wordscores* when the purpose of automated coding is to count valenced reasons in a text.

The sequential pattern of wordscores can be tied to reasons in two ways. The first is by counting hills and valleys using an empirically determined threshold value above or below which a hill or valley is counted. This is a conservative and more time consuming procedure but takes into account meaningful groupings of valenced terms. It is conservative because a list of features with few function words separating them will count as one reason rather than several. For example, if a person writes "better health, avoid disease, knowledge re future," that would produce one hill of positively valenced words even though three supportive

reasons are provided. Reasons stated as lists without the typical supplemental function words would produce an underestimate of the number of reasons.

A second approach makes an adjustment for the above possibility by separating the "positive reference text" and the "negative reference texts." Within each valence, wordscores of observed words are summed and divided by two.[7] Division by 2 reflects the empirical reality that a large number of entries function as bigrams, that is, pairs of words occurring with high frequency (Damerau, 1971).

The detailed analysis of bigrams is ignored and instead a simple heuristic of division by two takes its place. This heuristic is employed (1) because the selection of relevant bigrams is subjective and time-consuming, (2) the number of bigrams escalates exponentially even with a relatively small number of unique words, and (3) rules for selecting subsets of bigrams (e.g., top 10%) are arbitrary. The heuristic we employ is simple and straightforward even though it will require testing in a variety of other contexts before it can be adopted more generally. The formula for predicting the valenced reasons follows:

$$\begin{aligned}\text{Positive thoughts}_{\text{t}} &= \frac{\sum_{w}\text{positive wordscore}_{\text{wr}}}{2},\\ \text{Negative thoughts}_{\text{t}} &= -1 \times \frac{\sum_{w}\text{negative wordscore}_{\text{wr}}}{2}\end{aligned} \quad (2)$$

where "Positive thoughts$_t$" or "Negative thoughts$_t$" denotes the predicted number of positively or negatively valenced reasons in a target text *t*; "positive wordscore$_{wr}$" or "negative wordscore$_{wr}$" is a positive or negative wordscore of a word *w* in a target text *t*.

In the two exemplar responses mentioned above ("I think genetic tests will improve medical treatment and may help sick people" and "I think genetic tests will endanger privacy information and may discriminate against racial minorities"), the estimated computer coding would be about 2.35 (positive thoughts) and .03 (negative thoughts) in the positive response, and 1.93 (negative thoughts) and .12 (positive thoughts) in the negative response, which are close to the human coded number of reasons—two in both pro and con responses.

TEXTS FOR WORDSCORES AND AFFECTIVE INTONATION

Open-ended Responses: Argument Repertoire Questions

The textual data for testing comes from open-ended responses of a large sample of the general public about their positive and negative thoughts towards voluntary participation in genetics research and testing. Respondents were asked whether they had positive or negative views on an ethical issue in genetic testing by using the following question: "Would you say that you lean a little more toward "likely to volunteer" or a little more toward "unlikely" or can't you say for sure?").[8] If a person chose either "very likely" or "likely," then the person

[7]The integer of 2 changes the mean of estimated number of thoughts in a text, but does not change its variance because it is constant value. In order words, if researchers focus on the relationship between automated content analytic scores and other variables (e.g., correlation coefficient), then any constant number can be acceptable because it has no influence on variance and covariance structure.

[8]The proportion of the adamant neutral opinion holders is only 5% (n = 79 out of a total 1,961 respondents), and they did not receive any open-ended questions.

is defined as a pro; and a respondent is considered as a con if the respondent selected either "very unlikely" or "unlikely."

After identifying the issue stance of respondents, they were subsequently asked the reasons for their own views and anticipated reasons for why others might hold the opposite views. For example, one person whose issue position is positive regarding participation in genetics research would provide positive reasons to support his/her issue position, and would also anticipate negative reasons that might justify the opposing viewpoint. Specifically, respondents with positive views on voluntary participation in genetic tests are asked to answer two open-ended questions: (1) "What are the reasons you have for being in favor of volunteering to take a genetic test as part of a research study?" (i.e., Pros' own reasons), and (2) "What reasons do you think other people might have for being opposed to volunteering to take a genetic test as part of a research study?" (i.e., Pros' supposed reasons). Respondents with negative views on the issue receive two open-ended questions: (1) "What are the reasons you have for being opposed to volunteering to take a genetic test as part of a research study?" (i.e., Cons' own reasons), and (2) "What reasons do you think other people might have for being in favor of volunteering to take a genetic test as part of a research study?" (i.e., Cons' supposed reasons).

Two things should be emphasized. First, the valence of each response is clearly determined by the format of the open-ended questions (i.e., respondents themselves provided positive responses towards positive question and *vice versa*.) Second, responses vary in the number of thoughts provided to support one's own view or views of those with the opposite view.

The former outcome will be defined as "valence" (i.e., positive versus negative) and the latter as "valenced thoughts" (i.e., the number of positively valenced reasons and the number of negatively valenced reasons). In addition, two affectively[9] directional terms (i.e., both positivity and negativity) will be used to portray the issue position of a person's response. *Wordscores* is a method for predicting valence, and *Affective Intonation* is a method for estimating the number of valenced thoughts in a response.

Textual Data and Preprocessing Procedures

Survey respondents were drawn from a nationally representative panel maintained by Knowledge Networks, Inc., Menlo Park, California. The Knowledge Networks panel comprised a large number of households that have been selected through random digit dialing (RDD) and that agreed to accept free web TV equipment and service in exchange for completing periodic surveys online (AAPOR response rate II = 47%), indicating that open-ended responses were obtained online. The survey was conducted in March 2009 and had a total of 1,961 respondents. Of these, about 69% of respondents entered valid open-ended responses. The automated content analysis included a total of 2,786 responses (1,409 positive and 1,377 negative) generated from 1,435 respondents.

[9]The original developers of *Wordscores* (Laver et al., 2003) favor "(ideological) position" because their text samples are political documents.

Most studies using *Wordscores* do not preprocess the raw texts because the texts are carefully edited formal documents. However, our textual data requires some preprocessing because open-ended responses are informal and rife with errors. Overall, the preprocessing is not technically difficult and can be carried out after relatively simple training.

First, we corrected typos in the open-ended responses, to enhance the coding performance of computerized content analysis. We used popular word processing software (Microsoft Office®) to detect and correct typos and grammatical errors. To assure systematic procedures and simplicity, all typos and errors were replaced with the corrections that the word processing software recommended as the first choice. Second, symbols (e.g., @, $), punctuation marks (e.g., period, !, ?), and numerals (e.g., 100) were deleted in open-ended responses because those characters were rarely found and it seemed safe to assume that their wordscores might fluctuate around zero and therefore could be ignored without substantive consequences. Third, all upper-case words (e.g., DNA, FDA) were replaced by lower-case words (e.g., dna, fda) to keep consistency of the notation across survey respondents. Finally, we apply stemming procedures, such as replacing "companies" with "company," "is" with "be," and "increased" or "increasing" with "increase" in order to reduce irrelevant variance in reference texts. We used the textual mining software, 'tm' package in *R*[10] (Feinerer, 2008) to delete symbols, punctuation marks, and Arabic numbers; to convert upper-case words into lower-case words; and to unify stemmed words.

RESULTS

Does *Wordscores* Predict Valence of Informal and Nonpolitical Texts of the General Public?

Table 1 summarizes the descriptive statistics for reference texts as a function of their proportion of the total sample. As the size of the reference texts from the whole sample increases, the percent of "unique words (without redundant appearances) observed in reference texts" increases sharply before tapering off, while the percent of "total words (allowing redundant appearances) covered in reference texts" increases proportionally. For example, about 22% of reference texts (i.e., 600 responses out of 2,786 responses) covered a similar portion of total words (i.e., 21%) but almost 50% of unique words (see Table 1). What these data show is that a small subset of reference texts would be sufficient to scale most of words in the whole corpus for these samples of open-ended responses from a general population.

The left panel in Figure 2 presents the relationship between sample size (as a proportion) and the number of unknown responses in the target texts (as a percentage). After the sample reaches 22% of responses as reference texts, the percentage of unknown words drops to 1%, implying that a small subset is adequate to generate wordscores in the larger sample of target text.

[10]R is freely obtainable at www.r-project.org. The textual mining package ('tm') in R is available at http://cran.r-project.org/web/packages/tm/index.html

Panel two of Figure 2 (right hand side) shows that *Wordscores* distinguish positive responses from negative responses reasonably well.[11] Using only 7% of reference texts, *Wordscores* achieves acceptable reliability, that is, a Krippendorff's $\alpha = .61$ assuming nominal categories. As the proportion of reference texts increases, the reliability increases to a Krippendorff's $\alpha > .70$ (when 50% of responses are selected as reference texts).

These results indicate that *Wordscores* are reliable even with short, informal, and nonpolitical texts that are generated by a broad range of respondents. More importantly, even when the valence of only 30% of total responses is known, valence coding via *Wordscores* shows a satisfactory reliability level with nearly zero missing values (i.e., unknown).

Does *Affective Intonation* Predict Valenced Thoughts of Open-ended Responses?

Four scatterplots in Figure 3 show the relationship between human and machine coding of reasons using the heuristic procedures we have developed. A 45° line would indicate perfect association (dotted lines). The linear association across cases indicates substantial correspondence between manually coded thoughts and predicted thoughts via *Affective Intonation*. The highest reliability coefficient was .76 (Krippendorff's α assuming interval scaling) when predicting supposed positive thoughts (i.e., open-ended responses of negative opinion holders for why opponents think about the issue positively), and the lowest case was .56 when estimating own negative thoughts (i.e., open-ended responses of negative opinion holders for why they think of the issue negatively).

When the number of valenced reasons is small (fewer than 2), predicted number of reasons are slightly larger than those coded manually (i.e., solid lines are above dotted 45° lines). When the manual codes are above 3, the predicted codes are lower than the manual ones (i.e., dotted 45° lines are above solid lines).[12] In general, the mean of manually counted thoughts is comparable to the mean of predicted thoughts via *Affective Intonation*.

The results of Figure 3 show satisfactory coding for valenced thoughts in target texts. The comparisons in the figure are at the individual-response level for valenced thoughts.[13] Some automatized coding procedures only show successful prediction when data are aggregated into larger groups (e.g., if a person has multiple responses, the person's summed or mean

[11]Obviously, this finding is obtained via point estimate of a target text *t*'s valence without considering uncertainty of the text's point estimate (e.g., 95% of confidence interval). Since most open-ended responses are short, the point estimate of a text *t* has wider confidence interval, indicating that many open-ended responses are not clearly classified as "certainly negative" (i.e., the upper bound of 95% CI is less than 0) or as "certainly positive" (i.e., the lower bound of 95% CI is more than 0). For example, when 22% of open-ended responses are chosen as reference texts and the remaining 78% are treated as target texts (n = 2,186), 50% of target texts (n = 1,090) are classified as uncertain (i.e., 0 is located between the lower bound and the upper bound of 95% CI). However, out of the uncertain point estimates of target texts, 73% of automated coding (n = 799) based on point estimates are consistent with a human coder's valence classification, meaning that point estimates seems reasonable, despite the lack of the certainty mainly due to the short open-ended responses.

[12]Probably the slight over-interpretation of machine coding over manual coding whose scores are less than 2 might be caused by the summation function in our *Affective Intonation* because wordscores of semantically less determining words (e.g., people, information, in the previous example) are also included in the counted reasons although their values of positivity or negativity may not be substantial. The slight underestimation of machine coding over manual coding whose scores are above than 3 would show the limitation of our choice of linear rescaling approach (i.e., dividing the summed value by the integer 2) because a response containing many reasons is usually a list of one or two words, whose wordscores are in general less than 1, summarizing a separate thought like the previous example (i.e., "better health, avoid disease, knowledge re future"). Readers should be informed that these presumptions are not thoroughly examined via systematic comparison.

score across multiple responses can be aggregated at the individual-person level). Here we offer a conservative test of affective intonation at the individual level of prediction.

Predictive Validity of Computer Coding

The four scatterplots in Figure 3 show considerable correspondence between manual coding and computer coding using *Affective Intonation*. However, another criterion for validating the computer coding is to compare conclusions drawn about the computerized versus human coding in a predictive context. One simple test of "comparative predictive validity" is the relationship between the number of reasons and a respondent's cognitive ability as captured by educational achievement (Cappella et al., 2002). The results from this test are presented in Figure 4 and Table 2. The key to a test of "comparative predictive validity" is whether the inferences drawn about the predictor-outcome relationship are the same for the two versions of the data—in this case human and computerized.

Figure 4 shows means and their 95% confidence intervals for both methods of coding across three strata of educational achievement: (1) high school graduate or less, (2) some college education, and (3) bachelor's degree or higher. Consider first "own thoughts" (negative and positive). Educational achievement is significant for both positive and negative thoughts and there is no interaction effect. This implies that the pattern across education levels is the same for the two approaches.

There is a main effect between coding approaches that is indicative of higher scores for computerized techniques. However, both the pattern of results from low to moderate and moderate to high, and the significance tests between these levels, would be the same for the two types of coding. That is, the inferences drawn between across levels of education in terms of reasons provided would be identical in the two different modalities of coding. So this test for (valence of) own reasons indicates that the comparative predictive validity is the same.

Next, consider the pattern for "supposed thoughts" (positive and negative) across the two modes of coding. There is a significant main effect for education and no difference in the means for coding approach but a significant interaction for both positive and negative thoughts. The interaction effect is due mostly to an overestimation of reasons for the highest education group, especially for negative thoughts. However, the inferences that would be drawn across levels of education for supposed reasons would be the same despite the significant interaction. That is, low and moderate levels of education are no different in

[13]We also compare predictive validity of predicted thoughts against that of manually coded thoughts. For example, Cappella and colleagues (2002) examined a series of correlations between manually coded thoughts and established criteria variables, such as education level, issue knowledge, and patterns of media use or interpersonal talk. Theoretically, it is expected citizens who are more educated, more knowledgeable, and active discussants or media users will show more thoughts (i.e., reasons) towards an issue because they are politically sophisticated. If automated coding is comparable with manual coding, then a set of correlations between the number of thoughts and criteria variables should be similar, regardless of coding method (i.e., computers versus human coders). Thus, we carefully examined and compared two sets of correlations between the number of thoughts and criteria variables. Results show that two sets of correlations are indistinguishable and patterns of significance testing were highly consistent across all criteria variables. In sum, results demonstrate that correlations to various criteria for human and computerized codes achieve virtually the same level of validity. Like manually counted thoughts, predicted thoughts via Affective Intonation are well predicted by criteria variables, such as education, issue knowledge, and a respondent's information seeking behaviors.

supposed reasons while those with a bachelor's degree and above have more supposed reasons than either of the other educational groups.

The bottom line is that the predictive validity test (and many other similar tests with other predictors) yields inferences between the two approaches that are no different whether the coding is via human or computerized approaches.

SUMMARY AND DISCUSSION

In this study, *Wordscores* techniques developed in political science are introduced to communication researchers and adapted to the coding of valenced statements. Evidence is presented supporting *Wordscores* as a reliable, effective, and efficient content analytic method to predict opinion valence of ordinary open-ended, short, and informal textual responses. *Wordscores* succeeds in correctly predicting the valence of open-ended responses with satisfactory reliability (Krippendorff's $\alpha > .60$ at nominal level) compared to human coding, even when a relatively small proportion of responses is used for reference texts.

Affective Intonation is developed as a complementary technique for estimating the number of valenced thoughts or reasons contained in open-ended responses. Results with *Affective Intonation* demonstrate that the automated computer coding system provides a satisfactory level of reliability (Krippendorff's α ranges .56 to .76) that is comparable with manually counted valenced thoughts.

We believe that both *Wordscores* and *Affective Intonation* are useful methods to content analyze ordinary open-ended responses efficiently and consistently. First, our application of *Wordscores* in the context of valenced textual data is as reliable as original studies of conventional *Wordscores* where Pearson's $r \cong .71$ in 24 cases, ranging from .31 to .96 (see Klemmensen et al., 2007; Laver et al., 2003). Original applications were employed to predict the ideological orientation of formally prepared textual material and (usually) employed expert opinions about the texts' ideology. Our findings are encouraging because they imply that *Wordscores* is at least generalizable to text valence. Of course, replication of *Wordscores* in a variety of contexts is necessary because the tested domains of application are still limited.

Second, *Affective Intonation* is a straightforward and easy but reliable method to content-analyze valenced reasons in texts. Most automated content analytic methods focus on the successful prediction of the valence of a text but pay little attention to how strongly positive or negative the text is in its stated reasons.

Finally, our study adopts the assumption that words in a text are "data" (Benoit et al., 2009; Laver et al., 2003). This bottom-up (or empirical) approach has advantages. First, the approach is relatively easy to use because it ignores syntactic information. Although syntactic parsing (e.g., van Atteveldt, Kleinnijenhuis, & Ruigrok, 2008) provides useful information of semantic relationships between concepts (e.g., "Obama defeats McCain" is substantively different from "McCain defeats Obama"), in the contexts of "inferring valence of a text" or "counting the valenced thoughts" syntactic information may not contribute to the improvement of automated content analysis. Sophisticated syntactic parsers are

necessary in order to reliably extract the relational information based on syntax in a text. However, syntactic parsers frequently fail to work in informal or ungrammatical texts like open-ended responses or online texts, although the parsing performance has being improved continuously. Second, the bottom-up approach is flexible because it is unnecessary to use an established dictionary (Bantum & Owen, 2009; Pennebaker et al., 2001) or theoretically driven coding schemes (Laver & Garry, 2000; Schrodt & Gerner, 1994). For example, the Kansas Event Data System (Schrodt & Gerner, 1994) is a reliable coding system used to predict political change in the Middle East, Balkans, or West Africa. However, this system is not as useful in other contexts because it is highly specialized for international conflicts. However, both *Wordscores* and *Affective Intonation* are easy to use (available via popular statistical programs such as STATA or R[14]), and the choice of the affective dimension for machine coding depends only on researchers' interests.

The validity and reliability tests carried out here used data from a large scale survey of adults responding to questions about bioethical issues in genetics' research and testing and employing a methodology for open-ended questions that has been mainly used in political communication (Cappella et al., 2002). However this domain of testing does not limit the conclusions drawn about *Wordscores* or *Affective Intonation* to this context or method. These procedures can be readily applied to other domains of content analysis including thought-listing, "considerations," and other domains where the valence of textual documents (formal or informal) is of interest. In each of these potential contexts of application, the size of the data set to be coded is a factor. For example, a study using a single question about thought listing with a small sample of respondents or a content analysis of a few hundred news stories is probably best handled with procedures employing human coders. However, as the availability of large, inexpensive samples from on-line research companies has grown; as the availability of very large amounts of textual data from online and other sources has exploded, the need for more automated approaches that allow efficient and valid coding of large amounts of data is paramount.

We propose *Wordscores* and *Affective Intonation* as two automated content analytic methods useful for researchers in persuasion, politics, new media, content analysis, and public opinion who want to categorize large amounts of textual data quickly, systematically, and efficiently. We are not ready to recommend using *Wordscores* and *Affective Intonation* when a study's purpose is providing an absolute score for an aggregate population. This would include claims about the absolute number of reasons or the absolute degree of positivity (or negativity) in a sample of some type of text. However, we are confident that these procedures will provide valid claims across comparison levels (e.g., across levels of education, knowledge, experimental condition, types of stories) or in measures of association between output from *Wordscores* or *Affective Intonation* and some other criterion. In our own studies with large samples and multiple open-ended questions, *Wordscores* and *Affective Intonation* have replaced human coding.

[14]*Wordscores* package for STATA can be found at http://www.tcd.ie/Political_Science/wordscores/software.html, and R package (called *austin* by Willie Lowe) can be downloaded at http://www.williamlowe.net/software/

In this study, textual responses of survey respondents are clearly guided by the structurally valenced set of open-ended questionnaires focusing only one topic, which probably helps to improve the precision of *Wordscores* and *Affective Intonation*, despite many limitations (i.e., short, informal answers). However, in some contexts of content analysis, communication scholars have to confront more complex texts. In those contexts, coding performance of *Wordscores* and/or *Affective Intonation* could be less robust. We look forward to other tests of these procedures in other contexts and with other outcomes to build a base for the procedures' validity and to extend the magnitude of databases able to be explored in substantively important ways.

References

Bantum EOC, Owen JE. Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. Psychological Assessment. 2009; 21(1): 79–88. [PubMed: 19290768]

Benoit K, Laver M, Mikhaylov S. Treating words as data with error: Uncertainty in text statements of policy positions. American Journal of Political Science. 2009; 53:495–513.

Brock TC. Communication discrepancy and intent to persuade as determinants of counterargument production. Journal of Experimental Social Psychology. 1967; 3(3):296–309.

Cappella JN, Price V, Nir L. Argument repertoire as a reliable and valid measure of opinion quality: Electronic dialogue during campaign 2000. Political Communication. 2002; 19(1):73–93.

Damerau, FJ. Markov models and linguistic theory, an experimental study of a model for English. The Hague, The Netherlands: Mouton; 1971.

Duda, RO.; Hart, PE.; Stork, DG. Pattern classification. 2. New York. NY: John Wiley & Sons, Inc; 2000.

Fan DP, Holway WB. Media coverage of cocaine and its impact on usage patterns. International Journal of Public Opinion Research. 1994; 6(2):139–162.

Feinerer I, Hornik K, Meyer D. Text mining infrastructure in R. Journal of Statistical Software. 2008; 25(5):1–54.

Hopkins DJ, King G. A method of automated nonparametric content analysis for social science. American Journal of Political Science. 2009; 54(1):229–247.

Klemmensen R, Hobolt SB, Hansen ME. Estimating policy positions using political texts: An evaluation of the Wordscores approach. Electoral Studies. 2007; 26(4):746–755.

Krippendorff, K. Content analysis: An introduction to its methodology. Thousand Oaks, CA: Sage Publications; 2004.

Laver M, Benoit K, Garry J. Extracting policy positions from political texts using words as data. The American Political Science Review. 2003; 97(2):311–331.

Laver M, Garry J. Estimating policy positions from political texts. American Journal of Political Science. 2000; 44(3):619–634.

Lowe W. Understanding wordscores. Political Analysis. 2008; 16(4):356–371.

Martin LW, Vanberg G. A robust transformation procedure for interpreting political text. Political Analysis. 2008; 16(1):93–100.

Monroe BL, Schrodt PA. Introduction to the special issue: The statistical analysis of political text. Political Analysis. 2008; 16(4):351–355.

Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2008; 2(1):1–135.

Pennebaker, JW.; Francis, ME.; Booth, R. Linguistic inquiry and word count: LIWC 2001. Mahwah, NJ: Erlbaum; 2001.

Petty RE, Brinol P, Tormala ZL. Thought confidence as a determinant of persuasion: The self-validation hypothesis. Journal of Personality and Social Psychology. 2002; 82(5):722–741. [PubMed: 12003473]

Petty, RE.; Cacioppo, JT. Communication and persuasion: Central and peripheral routes to attitude change. New York, NY: Springer-Verlag; 1986.

Petty, RE.; Wegener, DT. Attitude change: Multiple roles for persuasion variables. In: Gilbert, DT.; Fiske, ST.; Lindzey, G., editors. Handbook of social psychology. Vol. 2. New York, NY: McGraw-Hill; 1998.

Popping, R. Computer-assisted text analysis. Thousand Oaks, CA: Sage Publications; 2000.

Price V, Cappella JN. Online deliberation and its influence: The electoral dialogue project in campaign 2000. IT and Society. 2002; 1(1):303–329.

Price V, David C, Goldthorpe B, Roth MM, Cappella JN. Locating the issue public: The multidimensional nature of engagement with health care reform. Political Behavior. 2006; 28(1): 33–63.

Price V, Neijens P. Opinion quality in public opinion research. International Journal of Public Opinion Research. 1997; 9(4):336–360.

Price V, Nir L, Cappella JN. Framing public discussion of gay civil unions. Public Opinion Quarterly. 2005; 69(2):179–212.

Romantan, A. A longitudinal model of social amplification of commercial aviation risks: Exploring United States news media attention to fatal accidents and media effects on air travel behavior, 1978–2001. Philadelphia, PA: University of Pennsylvania; 2004.

Schrodt PA, Gerner DJ. Validity assessment of a machine-coded event data set for the Middle East, 1982–92. American Journal of Political Science. 1994; 38(3):825–854.

Shah DV, Watts MD, Domke D, Fan DP. News framing and cueing of issue regimes: Explaining Clinton's public approval in spite of scandal. The Public Opinion Quarterly. 2002; 66(3):339–370.

van Atteveldt W, Kleinnijenhuis J, Ruigrok N. Parsing, semantic networks, and political authority using syntactic analysis to extract semantic relations from Dutch newspaper articles. Political Analysis. 2008; 16(4):428–446.

Yanovitzky I, Bennett C. Media attention, institutional response, and health behavior change. Communication Research. 1999; 26(4):429–453.

Yanovitzky I, Stryker JO. Mass media, social norms, and health promotion efforts. Communication Research. 2001; 28(2):208–239.

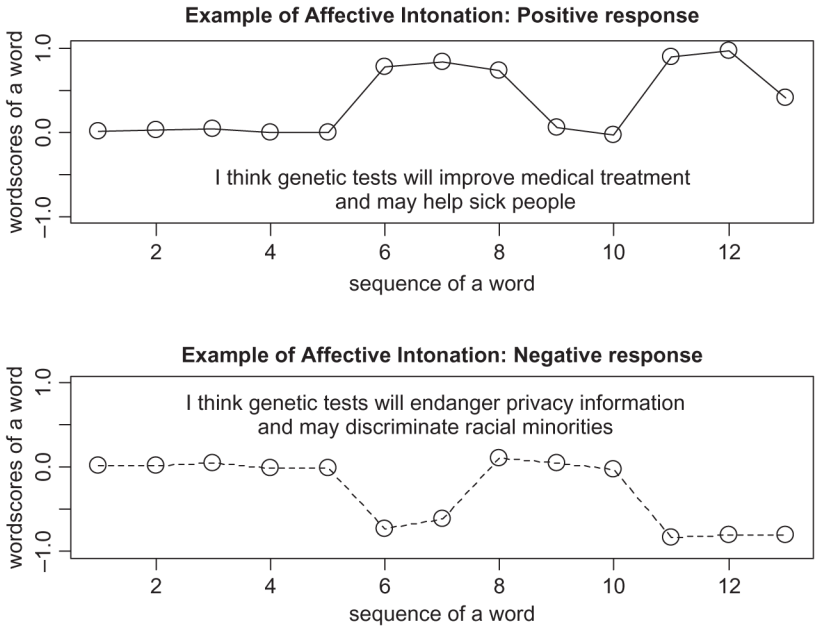

FIGURE 1. Affective Intonation

Change of Wordscores in Two Exemplar Sentences on the Word-sequence in a Response.

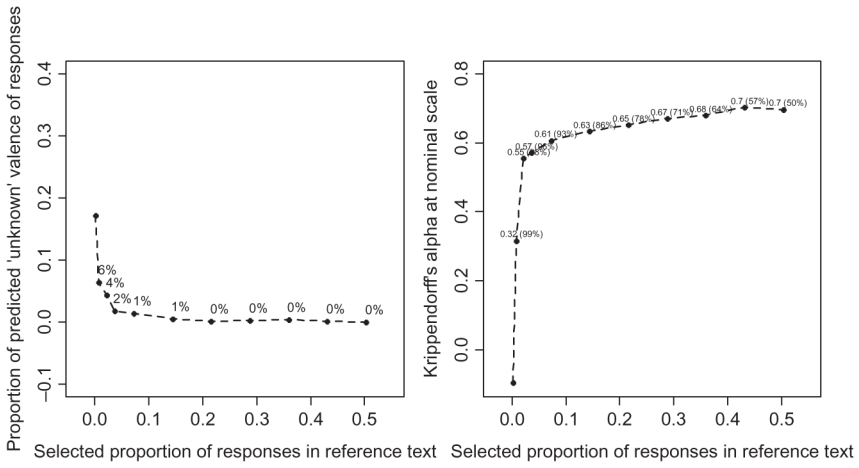

FIGURE 2.

Evaluation of Coding Performance of Wordscores According to the Selected Proportion of Reference Texts.

Note. In the left panel, percent of predicted 'unknown' valence of responses in target texts entered. In the right panel, values of Krippendorff's α between predicted valence of *Wordscores* and real valence of responses in target texts are displayed with the proportion of target texts out of the whole sample in parentheses. For example, Krippendorff's $\alpha = .32$ is a reliability coefficient between real valence (two categories: positive and negative) and predicted valence (three categories: positive, negative, and unknown) of responses in target texts, when only one percent of the whole sample was selected as reference texts, and the remained 99% of text sample was treated as target texts.

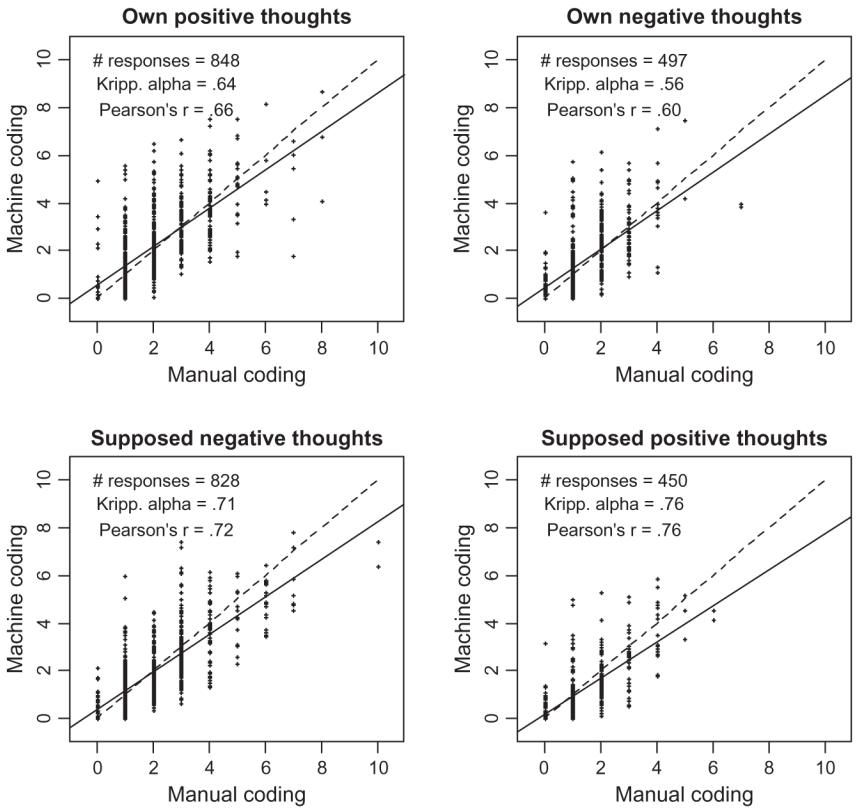

FIGURE 3.

Evaluation of Coding Performance of *Affective Intonation* Against Manually Coded Valenced Thoughts in Open-ended Responses.

Note. '# responses' denotes the included number of open-ended responses in four cases, and 'Kripp. alpha' is Krippendorff's reliability α at interval scale between manual and machine coding. "Own positive (or negative) thoughts" are open-ended responses containing thoughts why respondents hold a positive (or negative) issue position; and "Supposed positive (or negative) thoughts" are open-ended responses containing thoughts why the opposing side holds a positive (or negative) issue position. Dotted lines are 45° line, and solid lines represent estimated straight lines in each plot.

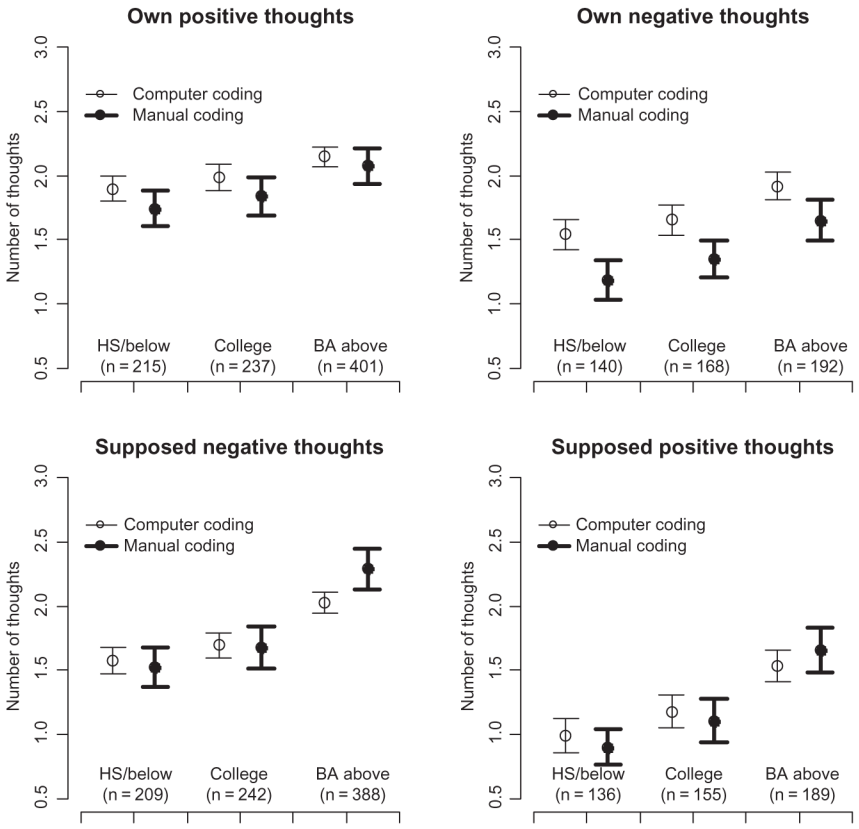

FIGURE 4.

Comparing the Effects of Educational Achievement on the Counted Number of Thoughts in Open-ended Responses via Computer Coding and Manual Coding.

Note. 'HS/below' presents respondents whose education level is high school graduate or less; 'College' characterizes people who graduate high school and attend some college; and 'BA above' stands for those who received a Bachelor or higher degree. The number of respondents in each category of educational achievement is entered in parentheses. "Own positive (or negative) thoughts" are open-ended responses containing thoughts why respondents hold a positive (or negative) issue position; and "Supposed positive (or negative) thoughts" are open-ended responses containing thoughts why the opposing side holds a positive (or negative) issue position.

TABLE 1

Descriptive Statistics of Reference Texts Randomly Sampled from the Whole Text Sample

Reference texts selected	Unique words	Total words	Average appearance of a word
4 responses (0.1 %)	42 (2%)	50 (0.15%)	1.19
20 responses (1 %)	119 (6%)	183 (1%)	1.54
60 responses (2 %)	291 (14%)	724 (2%)	2.49
100 responses (4 %)	373 (18%)	1, 049 (3%)	2.81
200 responses (7 %)	548 (26%)	2, 135 (6%)	3.90
400 responses (14 %)	854 (40%)	4, 727 (14%)	5.54
600 responses (22 %)	1, 042 (49%)	7, 168 (21%)	6.88
800 responses (29 %)	1, 199 (56%)	9, 388 (28%)	7.83
1,000 responses (36 %)	1, 330 (63%)	11, 920 (36%)	8.96
1,200 responses (43 %)	1, 430 (67%)	14, 009 (42%)	9.80
1,400 responses (50 %)	1, 562 (73%)	16, 637 (50%)	10.65
Total = 2,786 responses	Total = 2,127	Total = 33,473	15.74

Note. Total words allow redundant appearances of a word in reference text, but unique words do not allow such redundancy. For example, in a sentence “I have an apple and an orange,” the number of total words is seven, but the number of unique words is six because ‘an’ appears twice in the sentence. Average appearance of a word is the number of total words divided by the number of unique words.

TABLE 2

Testing Effects of Educational Achievement (between-subject factor) and Effect of Coding Method (computer vs. manual; within-subject factor) on the Counted Number of Thoughts in Open-ended Responses

	Own positive thoughts			Own negative thoughts		
	df	**F**	**partial η^2**	**df**	**F**	**partial η^2**
Educational achievement	2	7.869***	0.019	2	12.652***	0.051
Coding method (computer vs. manual)	1	13.532***	0.016	1	68.355***	0.127
Interaction (method*education)	2	0.750	0.002	2	0.457	0.002
Residual	813			470		

	Supposed negative thoughts			Supposed positive thoughts		
	df	**F**	**partial η^2**	**df**	**F**	**partial η^2**
Educational achievement	2	25.857***	0.060	2	21.109***	0.051
Coding method (computer vs. manual)	1	0.623	0.001	1	2.258	0.005
Interaction (method*education)	2	14.514***	0.034	2	8.459***	0.035
Residual	813			470		

Note.

*** $p < .001$. 'Educational achievement' is between-subject factor (i.e., one person has only one category), but 'coding method' is within-subject factor (i.e., one person has two observations because the person's response is 'manually' and 'automatically' coded).