Article

Sociological Methodology
1–30

DOI: 10.1177/00811750261444338
http://sm.sagepub.com

A Simulation-Based Slope Metric for Anchor List Reliability in Word Embedding Spaces

Marshall A. Taylor[1], Dustin S. Stoltz[2], Heather Harper[1], Sanuj Kumar[3], Sumanth Reddy Nandhikonda[3] and Luke Burks[4]

Abstract

Inducing semantic relations in word vector spaces and analyzing how other words or entire documents discursively engage these relations is a popular form of cultural analysis. The authors propose a reliability metric that is easily interpretable and agnostic to the type of relation. The metric, which the authors call the anchor reliability coefficient (*relco*), is found by creating an artificial document-term matrix of simulated documents that sequentially shift more of their tokens from relation-relevant anchor terms to nonanchor terms and then regressing the documents' similarity to an induced relation on the anchor inclusion score of the documents. The authors validate the metric at the word level with both expert- and crowdsourced dictionaries and at the document level with expert-annotated social media posts. The authors also provide some heuristic baselines for assessing reliability effect sizes and null hypothesis testing.

Keywords

word embeddings, reliability, simulation, semantic relations

Introduction

Inducing semantic relations, such as gender or affluence, in word vector spaces and analyzing how terms or entire documents discursively engage these relations is a popular form of cultural analysis (e.g., Arseniev-Koehler and Best 2025; Arseniev-Koehler and Foster 2022; Best and Arseniev-Koehler 2023; Daenekindt and Schaap 2022; Durrheim et al. 2022; Johnson 2024; Jones et al. 2020; Joseph and Morgan 2020; Kozlowski, Taddy, and Evans 2019; Nelson 2021; Pouliot and Patterson 2024; Taylor and Stoltz 2020a; Vann 2023; Voyer et al. 2022; Yoon and McCumber 2024). Defining these semantic relations involves selecting a set of "seed" or "anchor" terms.

[1]Department of Sociology, New Mexico State University, Las Cruces, NM, USA
[2]Department of Sociology and Anthropology, Lehigh University, Bethlehem, PA, USA
[3]Department of Computer Science, New Mexico State University, Las Cruces, NM, USA
[4]Department of Public Health Sciences, New Mexico State University, Las Cruces, NM, USA

Corresponding Author:
Marshall A. Taylor, New Mexico State University, Department of Sociology, 292B Science Hall, New Mexico State University, Las Cruces, NM 88003, USA.
Email: mtaylor2@nmsu.edu

However, research on how to assess the reliability of these anchor sets is somewhat nascent. Available tests are specific to the case of relations formed from juxtaposing antonym pairs–that is, inducing semantic directions or dimensions that have a bipolar structure, such as "gender" understood as "feminine" to "masculine." Yet one may wish to induce other types of relations that are not necessarily bipolar, such as engagement with "nature" (McCumber and Davis 2024).

We propose a reliability metric that is easily interpretable and agnostic to the type of relation, and which can support, not supplant, expert anchor list curation. The metric, which we call the anchor reliability coefficient (*relco*), is found by creating an artificial document-term matrix (DTM) of simulated documents that sequentially shift more of their tokens from relation-relevant anchor terms to nonanchor (e.g., randomly drawn, orthogonal, or conceptually distinct) terms and then regressing the documents' similarity to an induced relation on the anchor inclusion score of the documents.

We validate the metric in two ways that mirror how social scientists use relation induction: at the word level and the document level. For the word-level validation, we show that the proposed metric can distinguish between expert- and crowdsourced dictionaries that have been ranked along key semantic dimensions. These dictionaries come from moral foundations theory (Haidt and Graham 2007) and include words ranked as more or less associated with one of the five moral foundations (care, sanctity, fairness, loyalty, and authority). The first dictionary was created by experts in moral foundations theory, and the second was rated by U.S.-based nonexpert annotators. For the document-level validation, we show how adding words to an anchor list that have smaller contributions to the metric systematically decrease the predictive capacities of the induced relation–namely, the ability to correctly classify pandemic response posts in a corpus of tweets from 58 U.S. public health agencies. We also use simulations to generate suggested heuristic baselines for effect size comparison and null hypothesis testing.

Assessing the Reliability of Semantic Relations

Word Embeddings, Semantic Relations, and Anchor Lists

Word embeddings are vector representations of the meanings of words that summarize the contexts in which they occur. We do not go into detail here, as there are many useful introductions to word embeddings in the social sciences (e.g., Arseniev-Koehler 2022; Arseniev-Koehler and Foster 2022; Boutyline and Arseniev-Koehler 2025; Rodriguez and Spirling 2022; Stoltz and Taylor 2021). A key observation from this literature, though, is that these high-dimensional vector spaces encompass multiple meaningful "subspaces" (Boutyline and Arseniev-Koehler 2025; Stoltz, Combs, and Taylor 2023), which can be derived from the space using various techniques (Stoltz, Taylor, and Dudley 2024). These techniques include simple operations like subtraction or more complex ones like principal component analysis, and they can be used to define myriad semantic relations such as gender, morality, or political ideology.

For example, Best and Arseniev-Koehler (2023) analyze millions of news articles using embeddings to measure the stigma of more than 100 health conditions from 1980 to 2018. The news portrays a medical condition such as addiction as more immoral and disgusting than, say, dyslexia. But how are these relational concepts measured? Best and Arseniev-Koehler use a small list of "anchor" words to approximate the concept of disgust: *repulsive*, *disgusting*, and *gross*. The vectors for these terms are typically averaged to produce a single, new vector. Because this is a bipolar concept, they use anchor terms to approximate nondisgust: *appealing*, *captivating*, and *enticing*. Best and Arseniev-Koehler then measured how close a given condition's embedding vector is to either the stigmatizing or nonstigmatizing pole.

Validity and Reliability of Anchor Lists

Say an analyst is measuring how "gendered" a profession is by asking survey respondents how "feminine" or "masculine" the word *engineer* or *welder* is on a scale of 1 to 10. Assuming that "gender" is indeed an existing dimension in the culture of those being surveyed, we might ask whether this survey instrument accurately measures this actually existing "gender." If we, instead, define a gender relation from masculine to feminine in a word embedding space, and then measure the cosine similarity for the word *engineer* or *welder*, we must again consider whether this instrument accurately measures gender. Accuracy, in turn, will depend on the quality of the embeddings, the method used to derive the semantic relation, and, most important for our purposes, the words selected to anchor the relation.

In a review of 178 sets of anchor terms used in computer science and computational linguistics, Antoniak and Mimno (2021) find that researchers followed several strategies to select anchor sets. They often hand-curate the lists themselves, typically alongside close reading of source material; reuse anchor lists from other published sources; rely on lexical resources such as dictionaries, thesauri, or WordNet; crowdsource surveys using platforms such as Amazon Mechanical Turk; use natural language processing techniques to extract key terms from corpora; or use administrative data such as the U.S. census or political party roll-call votes.

More can be done to systematize anchor list construction, but most work in the social sciences that uses embeddings *does* discuss the process by which anchor words were selected (e.g., Best and Arseniev-Koehler 2023, online supplement; Jones et al. 2020:13–14; Taylor and Stoltz 2025, note 3). Kozlowski et al. (2019), for instance, include several anchor lists to define dozens of "cultural dimensions." To build these anchor lists, they rely on antonyms from five thesauri and the "subjective judgment on the part of the researcher" (Kozlowski et al. 2019:938). They then measure the correlations between human-rated associations of words (see also van Loon and Freese 2023), and those same words projected onto semantic directions created using different numbers of antonym pairs (they found an improvement as more pairs were added, but with diminishing returns). They also compare using random permutations of antonym pairings (they did not find much difference).

The literature on measurement in the social sciences provides two primary ways we should think about the accuracy of an instrument (Carmines and Zeller 1979). First, are we aiming at the right target when using a measure? This is typically called construct validity. Second, regardless of what is actually being measured, do the components of our instrument measure the same thing? This is called reliability. We focus on reliability. For example, for the feminine side of the gender relation, we might include *woman*, *women*, *female*, and *girl*—but should we include *queen* or *waitress*? The latter two are gendered terms, but they bring along additional semantic complexity related to occupation and status. This extra baggage may be a threat to reliability.

Beyond our own intuition, the intuition of others (expert or not), administrative data, or authoritative sources like thesauri, we have few formalized tools we can use to home in on our intended meanings. In what follows, we briefly review the reliability tools that do exist before proposing our more generalized addition. We note that these reliability tests should be treated as statistical heuristics: summary tools that are meant to *support*, not *supplant*, expert curation of good anchor lists that pass an "eyeball" test. In this sense, these reliability metrics, including the one we propose here, are akin to topic quality metrics in the topic modeling literature[1]: an additional tool in the analyst's toolkit for helping adjudicate between multiple possible upstream analytic decisions (i.e., choosing anchor words or a number of topics).

Prior Reliability Metrics

Each of the following metrics builds on the task of solving semantic analogies (Boutyline and Johnston 2025; Ethayarajh, Duvenaud, and Hirst 2019a; Mikolov, Yih, and Zweig 2013). The canonical example (from Mikolov, Yih, et al. 2013) of this is $\overrightarrow{queen} = \overrightarrow{king} + (\overrightarrow{woman} - \overrightarrow{man})$. In this example, subtracting $\overrightarrow{man}$ from $\overrightarrow{woman}$ is doing the heavy lifting by inducing a one-dimensional "gender" subspace from the otherwise high-dimensional embedding space. Because of the order of operations, this direction points toward "feminine" and away from "masculine." Thus, when adding $\overrightarrow{king}$ to this gender direction, it will move the vector toward the analogical equivalent of "feminine king." This approach is often called "offsetting," and it requires antonym pairs in our anchor set. Mikolov, Yih, et al. (2013) propose using this analogy task as one method to discover a reliable subspace (e.g., for gender). Specifically, this involves iterating through all the words in the embeddings to find the offset that would move $\overrightarrow{king}$ the closest to $\overrightarrow{queen}$.

Antoniak and Mimno (2021) offer some additional metrics to assess the reliability of the word set for pairs of antonyms. First, we could use principal component analysis to determine whether a single component adequately explains the variation of every offset (Bolukbasi et al. 2016; Ethayarajh, Duvenaud, and Hirst 2019b). Second, assuming that each word on one side of the direction should be maximally different from terms defining the opposing side, Antoniak and Mimno project each anchor onto the semantic direction created by the average offset. They refer to this as *coherence* (Boutyline and Johnston 2025 call this *rank separation*).

Boutyline and Johnston (2025:13) review these various reliability metrics and conclude by championing *parallelism:* the extent "the offsets between opposing anchors are parallel to one another." For instance, say we find $(\overrightarrow{woman} - \overrightarrow{man}) = \overrightarrow{feminine}_1$ and $(\overrightarrow{female} - \overrightarrow{male}) = \overrightarrow{feminine}_2$. We can then measure the similarity between these two induced relations, $sim\left(\overrightarrow{feminine}_1, \overrightarrow{feminine}_2\right)$. This similarity score tells us how closely we approximate the limiting case where $(\overrightarrow{woman} - \overrightarrow{man}) = (\overrightarrow{female} - \overrightarrow{male})$. If this similarity is high, we can conclude these two juxtaposing pairs of terms are parallel, and thus reliably measuring the same latent "gender" subspace.

Boutyline and Johnston (2025) demonstrate that parallelism is a well-supported measure of anchor term reliability. However, like previous metrics, it requires pairs of juxtaposing terms.[2] Therefore, it is best suited for instances in which the analyst wishes to derive a bipolar semantic direction from an embedding space. In what follows, we propose a more general, simulation-based measure of anchor list reliability.

A Simulation-Based Approach

Intuition

Let us introduce the logic of the relco with an intuitive example. Imagine a researcher studying moral foundations theory (Haidt and Graham 2007) wants to measure how words in their corpus engage with each of the five core moral foundations: care ↔ harm, sanctity ↔ degradation, fairness ↔ cheating, loyalty ↔ betrayal, and authority ↔ subversion. Specifically, they want to measure engagement with each virtue-coded pole (i.e., care, sanctity, fairness, loyalty, and authority). To do this, they create anchor lists of words indicating each virtue, then calculate the centroid (geometric center) of each list's word vectors in the embedding space. These

Table 1. Example Artificial Document-Term Matrix.

	w1	w2	w3	w4	w5	n1	n2	n3	n4	n5
doc 1	1	1	1	1	1	0	0	0	0	0
doc 2	0	1	1	1	1	1	0	0	0	0
doc 3	0	0	1	1	1	1	1	0	0	0
doc 4	0	0	0	1	1	1	1	1	0	0
doc 5	0	0	0	0	1	1	1	1	1	0
doc 6	0	0	0	0	0	1	1	1	1	1

Note: In the column headers, "w" represents "anchor word" and "n" represents "nonanchor word."

five centroids allow the researcher to measure (with cosine perhaps) how contextually similar any corpus word is to each virtue, where values closer to 1 indicate words that occur in similar discursive contexts.

This hypothetical analysis makes an important assumption: that the anchor lists are reliable, meaning any subcombination of words in the anchor list will not appreciably alter the centroid values. Assessing anchor list reliability empirically using the previously described methods (e.g., Boutyline and Johnston 2025) would not be applicable here, because the researcher is focusing only on the virtue anchor lists (meaning there are no vector offsets to derive).[3]

The researcher is using the Moral Foundations Dictionaries (Graham, Haidt, and Nosek 2009), a well-validated, expert-curated collection of words denoting each foundation's virtue and vice dimensions. But even for validated anchor lists, we need to assess reliability: do subsets of these words produce consistent centroid representations? We can use simulations to our advantage here. Using the "sanctity" anchor list as an example, we can randomly divide the 35 words[4] in the anchor list in half, using one half (plus one, given the odd number of words in the list) to define the centroid.

We then use the other half of the anchor list to initialize an empty (artificial) DTM (Stoltz and Taylor 2024:70–75). We next select another set of words of the same size—17, in this case—that are (1) randomly selected from other corpora or word lists that have no systematic relationship with the concept of sanctity, (2) systematically *not* related to the concept, or (3) conceptually distinct from the semantic relation of interest (e.g., one of the other sets of moral foundation virtue words). These words get appended to the DTM, resulting in a DTM with 34 columns and, as of now, 0 rows (documents). The first 17 columns are words from the anchor list; the last 17 columns are the nonanchor words.

As a final step, we add artificial "documents" as row vectors to the DTM. The first document equally distributes its tokens across the 17 anchor word columns. The final document equally distributes its tokens across the 17 nonanchor word columns. Each intermediate document sequentially shifts some of its tokens from an anchor word to a nonanchor word—transforming the first document into the last via sequential transitions. Table 1 illustrates this with a smaller example of five anchor words and five nonanchor words, with a single word token per word per document. For the "sanctity" anchor list specifically, we leverage a separate crowdsourced Moral Foundations dictionary of words collectively evaluated to *not* index sanctity discourse—a dictionary we discuss in more detail in the "Illustrations" and "Word-Level Validation" sections (Hopp et al. 2021).

Next, for each artificial document, we calculate concept engagement scores. We detail these scores in the next section, but for now it suffices to know that higher values indicate documents whose words are closer to the centroid and thus more engaged with the concept of "sanctity."

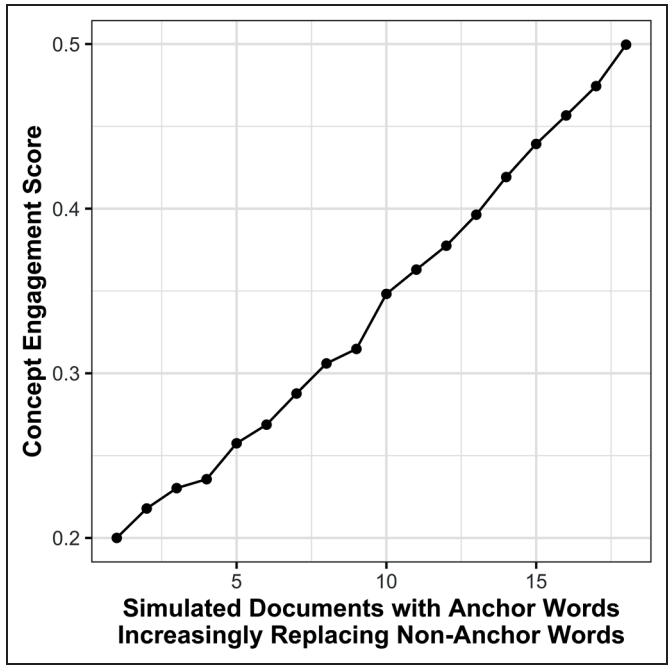

Figure 1. Relationship between concept engagement and anchor inclusion.
Note: Concept engagement scores derived from 17 random words in the "sanctity" anchor list/dictionary (Graham et al. 2009).

By design, the artificial documents are ordered by decreasing anchor content (increasing noise). Plotting each document's concept engagement score against its anchor inclusion reveals whether anchor words predict the concept better than nonanchor words do. Documents with more anchor words should engage the concept more than those with more nonanchor words. Figure 1 confirms this pattern for sanctity.

What does Figure 1 suggest about anchor list reliability? Put simply, anchor terms *not* used to form the centroid predict the centroid better than nonanchor words do. The anchor terms thus index the same thing—whatever that may be[5]—better than we would expect if the anchor list was a collection of other, nonanchor words. Importantly, analysts can choose whether nonanchor words are randomly drawn, orthogonal to the anchor words, or conceptually distinct—the metric accommodates all three approaches. (We return to this point later.)

This is the basic intuition the relco tries to quantify. This example works with only one random partition of an anchor list and with the anchor terms in the simulated DTM in only one order. We now turn to a more in-depth treatment of the relco itself and how we address these issues—as well as questions of the statistical significance of the coefficient.

The Anchor Reliability Coefficient

In this section, we outline our bootstrapped simulation-based regression method for quantifying anchor list reliability using the bivariate relationships between concept engagement and anchor inclusion scores. We also consider the metric's suitability for hypothesis testing. The method for deriving the coefficient is summarized with a schematic in Figure 2.

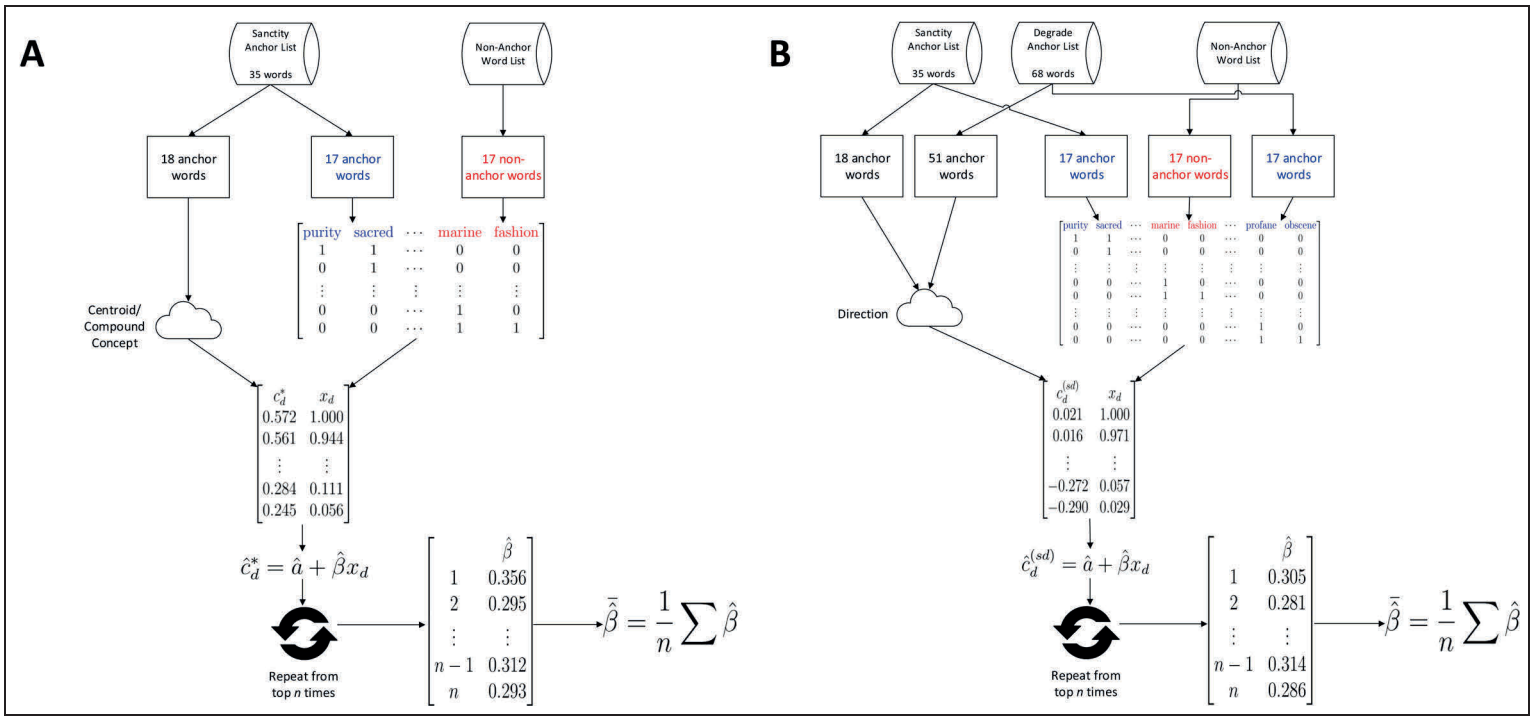

Figure 2. Metric schematic.

Note: (A) Metric workflow with semantic centroids and compound concepts. (B) Metric workflow with semantic directions. The "sanctity" anchor lists are used here for illustration purposes only; the workflow is agnostic to the anchor list. The nonanchor words could be randomly drawn words, orthogonal words, or words considered conceptually distinct from the semantic relation being indexed by the anchor words; it is at the analyst's discretion.

Finally, we present a method for derived word-level reliability scores, that is, each anchor word's contribution to the global relco.

Concept Engagement. The relco first requires a document-level measure of engagement with the semantic relation of interest. We use concept mover's distance (CMD) scores (Stoltz and Taylor 2019; Taylor and Stoltz 2020a, 2020b). In the "intuition" example above, CMD measures the extent to which each of the documents engages the concept of "sanctity" based on the minimum cost necessary to transform the words in each document into another "pseudodocument" (Stoltz and Taylor 2019:297–99) that consists only of a single token representing the centroid we defined using the other half of the anchor list. We do not cover the mathematical details for CMD here. We instead take the CMD scores themselves as our starting point, refer readers interested in such details to Appendix A and provide a brief overview of what CMD scores represent below.

CMD scores quantify, within the (typically) [0,1] interval when unstandardized,[6] the extent to which a document engages the semantic relation of interest. These semantic relations can be *compound concepts*, *semantic centroids*, or *semantic directions*. Compound concepts are represented as a dictionary of anchor list words but not reduced down to a single centroid (for more details on compound concepts, see Stoltz and Taylor 2019). Semantic centroids are represented as single vectors in the embedding space, computed as the arithmetic mean of the word vectors in the anchor list; CMD scores with compound concepts and semantic centroids have similar interpretations, where higher scores indicate a document that engages more the concept represented by the anchor list. A semantic direction, on the other hand, is a vector representation of antonymic anchor lists, that is, with one set representing one pole of a relation (e.g., "feminine") and the other set representing the other pole (e.g., "masculine"). CMD scores with semantic directions therefore have a polar interpretation: higher scores represent a document that engages more with pole 1 and lower scores represent a document that engages more with pole 2.

We will refer to the CMD score for document d on compound concepts as $c_d^{(cc)}$, where d is a document in d-by-t DTM $\mathbf{M}$ and t is a word. Similarly, we will refer to the CMD score for d on semantic centroids and semantic directions as $c_d^{(sc)}$ and $c_d^{(sd)}$, respectively. Lastly, we will let $\boldsymbol{\Psi}$ be our anchor list.

A Statistical Model of Reliability. Focusing on compound concepts and semantic centroids for a moment, let us assume that, before we derive our concept engagement vector $\mathbf{c}$, we subset anchor list $\boldsymbol{\Psi}$ to only include a random *half* of the words in the list (in the case of an odd-numbered anchor list, we add the extra word to the anchor list; we assume an even-numbered anchor list for the remainder of this section for clarity). Let the random subset that we will use to define the concept be $\boldsymbol{\Psi}_y$ and the remaining subset that we *will not* use to define the concept be $\boldsymbol{\Psi}_x$. Now assume that DTM $\mathbf{M}$ is not derived from a real corpus, but instead is an artificial DTM like we discussed earlier. The first k columns of this artificial $\mathbf{M}$ are the words in $\boldsymbol{\Psi}_x$. We then take another list of words that are either randomly drawn from a corpus or corpora, known to be orthogonal to the semantic relation of interest, or index a conceptually distinct semantic relation (e.g., an anchor list for a *different* centroid or compound concept). Let this word list be $\boldsymbol{\Upsilon}$, which is the same length (k) as $\boldsymbol{\Psi}_x$. We let the next k columns in artificial $\mathbf{M}$ be the words in $\boldsymbol{\Upsilon}$.

We then populate $\mathbf{M}$ with binary "document" row vectors in exactly the same way as depicted in Table 1, such that each (d, t) cell is populated as

$$M_{1t} = \begin{cases} 1 & \text{if} \quad t \in \boldsymbol{\Psi}_{\boldsymbol{x}} \\ 0 & \text{if} \quad t \in \boldsymbol{\Upsilon} \end{cases} \tag{1a}$$

$$\vdots$$

$$M_{(k+1)t} = \begin{cases} 0 & \text{if} \quad t \in \boldsymbol{\Psi}_{\boldsymbol{x}} \\ 1 & \text{if} \quad t \in \boldsymbol{\Upsilon} \end{cases}. \tag{1b}$$

Note from equation (1b) that the number of rows in **M** must be equal to $k+1$ if we sequentially shift one and only one $\boldsymbol{\Psi}_x$ cell count (a frequency of 1) to one $\boldsymbol{\Upsilon}$ cell count. **M**, then, will have dimensions $(k+1) \times (k+k)$. We then use the other anchor words in $\boldsymbol{\Psi}_y$ to derive the concept engagement scores for either a compound concept or a semantic centroid.

Equation (1) requires a small modification in the case of semantic directions. In this case, anchor lists consist of *two* sets of words, one defining the positive pole and one defining the negative pole. Simulated DTM **M** needs to be of dimensions $(k+k+1) \times (k+k+k)$. We populate this version of **M** with

$$M_{1t} = \begin{cases} 1 & \text{if} \quad t \in \boldsymbol{\Psi}_{+,\boldsymbol{x}} \\ 0 & \text{if} \quad t \in \boldsymbol{\Upsilon}, \boldsymbol{\Psi}_{-,\mathbf{x}} \end{cases} \tag{2a}$$

$$\vdots$$

$$M_{(k+1)t} = \begin{cases} 0 & \text{if} \quad t \in \boldsymbol{\Psi}_{+,\boldsymbol{x}}, \boldsymbol{\Psi}_{-,\mathbf{x}} \\ 1 & \text{if} \quad t \in \boldsymbol{\Upsilon} \end{cases} \tag{2b}$$

$$\vdots$$

$$M_{(k+k+1)t} = \begin{cases} 0 & \text{if} \quad t \in \boldsymbol{\Upsilon}, \boldsymbol{\Psi}_{+,\mathbf{x}} \\ 1 & \text{if} \quad t \in \boldsymbol{\Psi}_{-,\boldsymbol{x}} \end{cases}. \tag{2c}$$

where positive and negative anchor sets $\boldsymbol{\Psi}_+$ and $\boldsymbol{\Psi}_-$ are randomly split into x and y subsets, just as we did with compound concepts and centroids. We use the remaining anchor words $\boldsymbol{\Psi}_{+,y}$ and $\boldsymbol{\Psi}_{-,y}$ to derive the semantic direction concept engagement scores.

Finally, we can leverage the known row order of **M**—of either the equation (1) or equation (2) variety—to define a statistical model to assess how predictive the left-out anchor words are of concept engagement (for any three relation types) relative to random/orthogonal/distinct words using simple bivariate linear regression:

$$\hat{c}_d = \hat{a} + \hat{\beta} x_d, \tag{3}$$

where $\hat{a}$ is the standard intercept estimate, $\hat{\beta}$ is the slope estimate indicating the change in mean concept engagement for each additional anchor inclusion score increase, and $\hat{c}_d$ is solved using the standard ordinary least squares estimator. The $\hat{\beta}$ term is the relco for this one particular anchor list partition and with the anchor and nonanchor words in this particular order in the DTM. The x_d is the anchor inclusion score of the dth document, which is scaled by the vector's L_∞-norm to prevent $\hat{\beta}$ from being upwardly biased for smaller anchor lists. Scaling by the L_∞-norm involves dividing each anchor inclusion by the largest value so the values are restricted within the [0,1] interval with 1 equal to the highest anchor inclusion.[7]

To protect against reporting a relco that is biased by either of these forms of randomized subsetting, we add two further steps to the calculation. First, we simulate many runs that each start with a random partition of the anchor list (or anchor *lists*, in the case of a direction) without replacement. Let the number of runs be n. Second, for each n, we randomize the order of the k anchor terms and the k nonanchor terms (but always letting the anchor terms be the first k columns and the nonanchor terms always being the second k columns, and in the case of directions, letting the positive anchor terms be the first k columns, the nonanchor terms always be the second k columns, and the negative anchor terms be the last k columns).

The n runs result in a $1 \times n$ vector $\hat{\boldsymbol{\beta}}$. The final relco is the mean of this vector:

$$\bar{\hat{\beta}} = \frac{1}{n}\sum \hat{\beta}. \tag{4}$$

In the case of compound concepts and semantic centroids, $\bar{\hat{\beta}}$ can be interpreted as the average estimated change in mean concept engagement for each additional anchor inclusion score increase. Larger positive values indicate an anchor list that is more positively predictive of itself than a list of random words, meaning that the words in the anchor list "hang together" more so than we would expect if the anchor list was a random collection of words. In the case of semantic directions, $\bar{\hat{\beta}}$ can be interpreted as the average estimated change in mean positive pole engagement for each additional positive anchor inclusion score increase. As with compound concepts and semantic centroids, larger positive values indicate antonymic anchor lists that both (1) hang together more than we would expect if they were random collections of words and (2) are negatively predictive of one another (meaning that positive words predict positive words and negative words predict negative words).

We can leverage the fact that each n run starts with a random partition of an anchor list to make plausible assumptions about the sampling distribution of $\bar{\hat{\beta}}$. Each $\hat{\beta}$ in $\hat{\boldsymbol{\beta}}$ is generated from an $\mathbf{M}$ where $\mathbf{M}^1, \mathbf{M}^2, \mathbf{M}^3, \cdots, \mathbf{M}^{\mathbf{n}} \overset{iid}{\sim} \mathcal{U}(\Omega)$ (and $\mathcal{U}(\Omega)$ is the uniform probability distribution across the sample space of all possible $\mathbf{M}^{\mathbf{i}}$), accounting for every possible random partition (without replacement) of $\boldsymbol{\Psi}$ and every possible permutation of $\mathbf{M}$ column word order across the $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Upsilon}$ lists.[8] We can therefore understand the $1 \times n$ $\hat{\boldsymbol{\beta}}$ vector as a random sample of size n of relcos and $\bar{\hat{\beta}}$ as a sample estimate of μ, the true and unknown mean relco. Under central limit theorem assumptions, we know, given a sufficiently large n, that $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\mu_{\bar{\hat{\beta}}}, \frac{\sigma}{\sqrt{n}}\right)$ and $\mu_{\bar{\hat{\beta}}} \approx \mu$ ($\mu_{\bar{\hat{\beta}}}$ is the mean of the sampling distribution of $\bar{\hat{\beta}}$). Using an estimate of the standard error, $\frac{s}{\sqrt{n-1}}$ (where s is the observed standard deviation of vector $\hat{\boldsymbol{\beta}}$), we can generate confidence intervals and carry out standard hypothesis tests of μ, including, importantly, right-tailed one-sample tests of the null that μ is equal to or less than 0 or some other possible parameter the researcher deems to be low enough to indicate poor anchor list reliability (we return to the topic of baselines in a later section).

Word Contributions to $\bar{\hat{\beta}}$. The $\bar{\hat{\beta}}$ term is a "global" measure of anchor list reliability. What is missing is an analysis of which words in the anchor list contribute to high or low values of $\bar{\hat{\beta}}$. We can derive these contributions using recent insights into the two-mode network structure of regression output. As Schoon, Melamed, and Breiger (2024) point out, the slope coefficient estimate $\hat{\beta}$ from a linear regression is a duality (Breiger 1974; Mohr and Duquenne 1997): it

can be expressed as a function of the two variables it links as well as a weighted function of the observations in the data.

Consider an individual run of equation (3) to generate 1 of n relcos. Let this coefficient be $\hat{\beta}_1$. For compound concepts and semantic centroids, we can express each simulated document's contribution to the fully standardized version of $\hat{\beta}_1$, $z(\hat{\beta}_1)$, as

$$z(\hat{\beta}_1) = \sum_{d=1}^{k+1} \frac{z(x_d)}{\sum_{d=1}^{k+1} z(x)^2} z\left(c_d^{*}\right) = \sum_{d=1}^{k+1} z(\phi), \tag{5}$$

where the * indicates applicability for both compound concepts and centroids. And for semantic directions,

$$z(\hat{\beta}_1) = \sum_{d=1}^{k+k+1} \frac{z(x_d)}{\sum_{d=1}^{k+k+2} z(x)^2} z\left(c_d^{(sd)}\right) = \sum_{d=1}^{k+k+1} z(\phi). \tag{6}$$

We can then convert $z(\phi_d)$ back into the metric unit contributions for all three relation types with

$$\phi_d = z(\phi_d) \frac{sd(c_d)}{sd(x_d)}. \tag{7}$$

The ϕ_d term tells us the contribution of d to $\hat{\beta}_1$. This is the case for each of the n runs, so $\sum_{d=1} \phi_1 = \hat{\beta}_1$, $\sum_{d=1} \phi_2 = \hat{\beta}_2$, $\sum_{d=1} \phi_3 = \hat{\beta}_3$, $\cdots$, $\sum_{d=1} \phi_n = \hat{\beta}_n$.

The task then becomes how to turn each per-run simulated document's contribution to the per-run relco into each anchor word's contribution ($\eta_{\Psi_k 1}$), where, as before, Ψ_k is a word in anchor list $\boldsymbol{\Psi}$. We can take advantage of the fact that each d in each simulated $\mathbf{M}$ is sorted by the anchor inclusion score. The first simulated document, d_1, has an anchor inclusion score of 1, meaning that it consists of only ones in the k anchor word columns and only zeros in the k nonanchor word columns. As such, the difference in the document contribution to any $\hat{\beta}$ between d_1 and d_2 tells us the change in contribution when the first anchor word (M_{11}) is moved to the first nonanchor word ($M_{1[k+1]}$). We call this quantity $\eta_{\Psi_k 1}$, which is the change in contribution to $\hat{\beta}_1$ when anchor word Ψ_k is replaced by nonanchor word $\boldsymbol{\Upsilon}_k$ (when Ψ_k is the first word in that particular $\mathbf{M}$). We can write this as

$$\eta_{\Psi_k} = \phi_{d_1} - \phi_{d_2} | name(M_{,1}) = \Psi_k, \tag{8}$$

where name($M_{,1}$) is the name (word) for the $M_{,1}$ column. For semantic directions, we find η for positive word $\Psi_{+,k}$ the same way. To find η for negative word $\Psi_{-,k}$, we simply subtract the contribution from the $k+k+1$ document—the document with all nonanchor words save for one negative anchor word—from the contribution from the $k+k$ document, the document consisting of only nonanchor words. In the case of directions, we have

$$\eta_{\Psi_{+,k}} = \phi_{d_1} - \phi_{d_2} | name(M_{,1}) = \Psi_{+,k} \tag{9a}$$

$$\eta_{\Psi_{-,k}} = \phi_{d_{k+k}} - \phi_{d_{k+k+1}} | name(M_{,k+k+1}) = \Psi_{-,k}. \tag{9b}$$

As with the relcos, we want to bootstrap $\boldsymbol{\Psi}$ and repeat the calculations n times and report the mean contribution for each anchor word Ψ_k:

$$\bar{\eta}_{\Psi_k} = \frac{1}{n}\sum \eta_{\Psi_k} \tag{10a}$$

$$\bar{\eta}_{\Psi_{+,k}} = \frac{1}{n}\sum \eta_{\Psi_{+,k}} \tag{10b}$$

$$\bar{\eta}_{\Psi_{-,k}} = \frac{1}{n}\sum \eta_{\Psi_{-,k}}. \tag{10c}$$

With sufficient n relative to the length of the anchor set, each anchor word should have a reportable mean contribution. To ensure each anchor word is positioned in M_{11} (or, in the case of the negative pole of a semantic direction, $M_{[k+k][k+k+1]}$) across a reasonably small number of runs, we assign each anchor word in the first of the n runs with the same probability of selection into that column position and then reduce the probability for that word to be selected in that position in the following run by a penalty factor of 0.5. The probability of the word is further reduced by this same factor after each run where it was selected.[9]

For compound concepts, semantic centroids, and the positive pole of a semantic direction, larger values of $\bar{\eta}_{\Psi_k}$ indicate that, on average, any given $\hat{\beta}$ shrinks more when the frequency for that anchor word is replaced with the equivalent frequency of a nonanchor word. For the negative pole of a semantic direction, smaller values (closer to -1) indicate that, on average, any given $\hat{\beta}$ shrinks (but grows in absolute value) when the frequency for a nonanchor word is replaced with the equivalent frequency of that negative anchor word. We can therefore think of these values as word contributions to the global relco, $\overline{\hat{\beta}}$ (although they will not sum to be equal to $\overline{\hat{\beta}}$).

These word contributions are generated from the same simulated DTMs randomly drawn from a uniform probability distribution. We can make the same assumptions about the sampling distribution of each Ψ_k's $\bar{\eta}$ as we do for the sampling distribution for $\overline{\hat{\beta}}$: namely, assuming that $\boldsymbol{\eta}_{\Psi_k}$ is a $1 \times n$ vector of η contributions from anchor word Ψ_k, then $\boldsymbol{\eta}_{\Psi_k} \sim \mathcal{N}\left(\mu_{\bar{\eta}_{\Psi_k}}, \frac{\sigma_{\bar{\eta}_{\Psi_k}}}{\sqrt{n}}\right)$. This means, as with $\overline{\hat{\beta}}$, we can generate confidence intervals and conduct relevant hypothesis tests with each anchor word's contribution. This can come in handy for determining which words to mark as candidates for anchor list removal, as we will point out shortly.

Illustrations

We now illustrate relco and its word contributions with three applications: to assess the reliability of anchor words relative to (1) orthogonal words, (2) random words, and (3) conceptually distinct words. Each illustration consists of compound concept and semantic direction examples.

We will stick with the sanctity anchor list. Do the 35 words in the sanctity anchor list seem to measure the same thing relative to orthogonal, random, and conceptually distinct words? And do we increase the extent to which the words measure the same thing if we prune out anchor words that do not "hang together" with the other anchor words?

To answer this question, we first create three sets of 1,000 (n in the section above) simulated DTMs, each time randomly partitioning the anchor list into 18 words to define the compound concept and the remaining 17 as the first 17 columns in the DTM. Then, for each of the three sets, we add 17 words as the last 17 columns in each DTM. The set of 17 words are from the orthogonal, random, and conceptually distinct lists.

For the orthogonal list, we return to the crowdsourced extended Moral Foundations Dictionary, in which a collection of more than 3,200 English words were rated by U.S.-based annotators to indicate the extent to which that word is associated with each moral foundation (Hopp et al. 2021). Each word received a weight, which is the probability the word was annotated as an instance of that moral foundation. We took the 17 words with the lowest probabilities on the sanctity ↔ degrade foundation. We then populated each simulated DTM following equation (1). For the random list, we randomly drew 17 words in the Norvig Google Trillion Word Corpus (Norvig 2009) that occur more than 3,000 times and do not include the anchor list terms. Finally, for the conceptually distinct list, we randomly drew 17 words in the "authority" moral foundation dictionary, excluding any words also in the sanctity list.

Then, for each of the 1,000 simulated DTMs per set, we randomly shuffle (without replacement) the 17 anchor words across columns 1 through 17 and then randomly shuffle (without replacement) the 17 nonanchor words across columns 18 through 34. We then calculate the concept engagement score for each of the documents in each of the 1,000 DTMs per set and regress the score on the anchor inclusion score. The mean across all 1,000 slope estimates is the global relco. This resulted in three global relcos, one each assessing the sanctity list relative to orthogonal words, random words, and conceptually distinct words. We also get the contributions of each of the 35 anchor words per each set.

The left panels of Figure 3 depict each of the 1,000 slope estimates and the relco, the mean slope (the blue line), for each of the three sets. The box plots show the distribution of the 1,000 concept engagement scores per each of the 18 anchor inclusion score positions, where higher values indicate higher anchor inclusion.

The global relcos were 0.301, 0.320, and 0.261, all of which are positive. This suggests the anchor terms are more predictive of one another than are nonanchor words—of the orthogonal, random, and conceptually distinct varieties. But are these values statistically distinguishable from zero? A series of right-tailed one-sample t tests of H_0: $\mu \leq 0$ suggests that we can, in fact, confidently state that each of the true global relcos observable if we had managed to draw all possible simulated DTMs from the sample spaces are likely not equal to or less than zero (t = 235.52, 274.06, and 218.68). (A null hypothesis of zero might not be conservative enough of a test for reliability. We suggest some alternative null hypothesis baselines in the "Baselines" section.)

We can similarly assess the 35 words in the anchor list. We will focus on the anchor word contributions relative to the orthogonal nonanchor words for the sake of brevity. Table 2 lists the words in descending order of contribution, along with their 95 percent confidence intervals (two-tailed). *Saint*, *saints*, *saintly*, *churches*, and *church* have among the largest mean contributions; for example, on average, a simulated document's contribution to the relco decreases by an average of 0.014 when *saint*, as the first anchor word in the DTM, is converted into a nonanchor word. *Upright* has the smallest contribution but is still statistically significant. All the words have statistically significant positive contributions, in fact. If we *did* have a $\bar{\eta}$ that was statistically indistinguishable from zero, that word might be a candidate for removal from the anchor list. In this case, all 35 $\bar{\eta}$ being statistically significant, with and without multiple-comparison adjustments, might signal that this expert-curated dictionary is quite reliable.

As a semantic direction, we have two anchor lists for the sanctity ↔ degrade moral foundation: the list of 35 sanctity words we have been working with, along with a list of 54 degrade words (e.g., *profane*, *stain*, *exploit*, *obscene*). For this type of semantic relation, we first randomly split the sanctity word list into separate lists of 18 words to define the positive pole of the direction and the other 17 for the DTM, as before. We also randomly split the degrade list into two separate lists, one set of 37 words to define the negative pole of the direction and

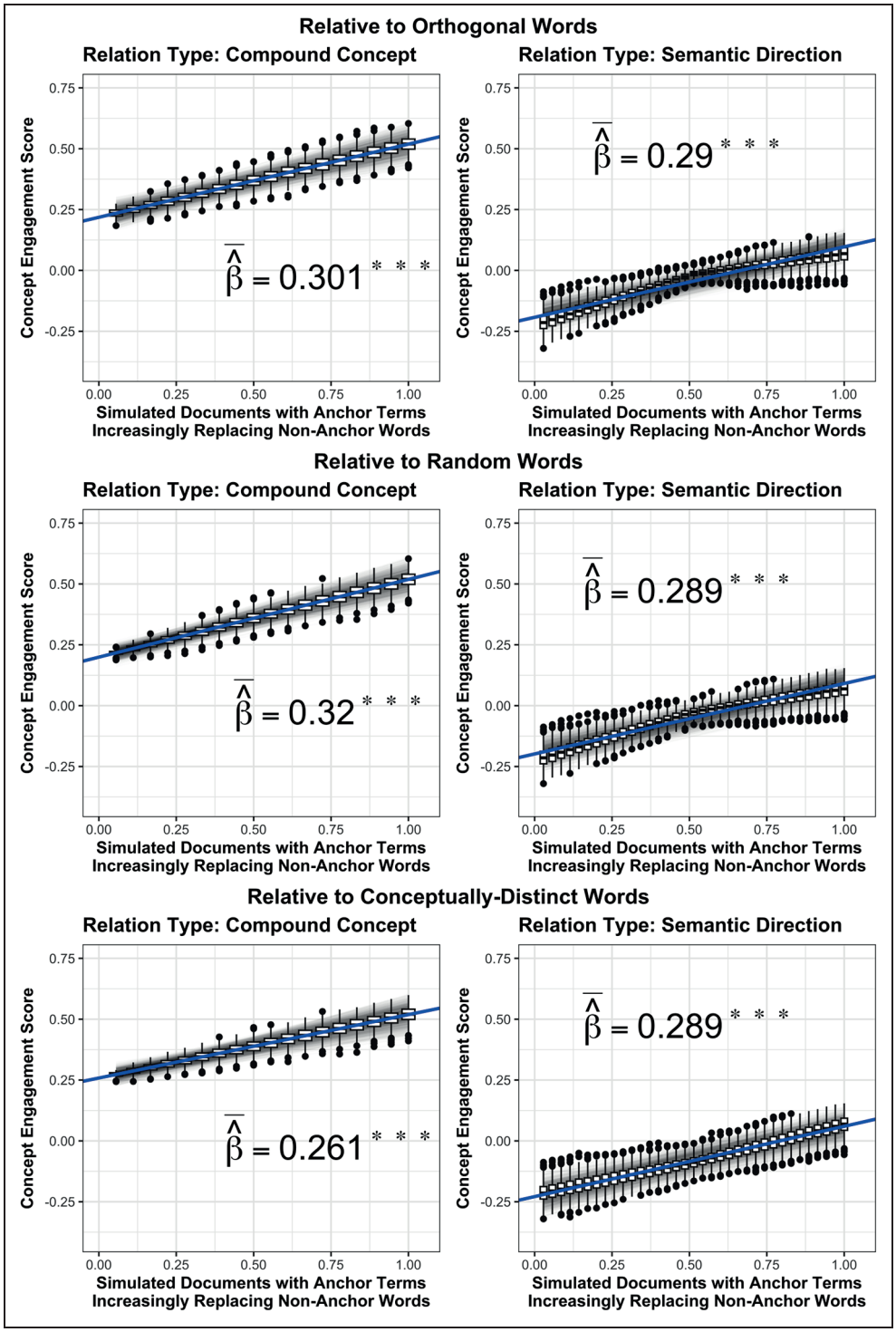

Figure 3. Sanctity anchor list reliability relation.

Note: Results from 1,000 runs. The anchor inclusion score, the *x*-axis, is normalized with the L_∞-norm. The embeddings are the pretrained English fastText embeddings (Mikolov et al. 2018).

***$p < .001$ (right-tailed).

Table 2. Word Contributions for Compound Concept Anchor List.

Anchor Word (Ψ_k) (Ψ_k)	$\bar{\eta}\times 100$	95% CI
saint	1.368	1.277–1.458
saints	1.238	1.182–1.384
saintly	1.189	1.076–1.302
churches	1.184	1.011–1.356
church	1.163	1.024–1.302
	⋮	
sterile	0.794	0.716–0.871
maiden	0.728	0.660–0.797
innocent	0.701	0.615–0.787
refined	0.661	0.587–0.735
upright	0.484	0.416–0.553

Note: Values are scaled by a factor of 100 for presentability. CI = confidence interval.

Table 3. Word Contributions for Semantic Direction Anchor List.

Anchor Word ($\Psi_{+,k}$)	$\bar{\eta}\times 100$	95% CI	Anchor Word ($\Psi_{-,k}$)	$\bar{\eta}\times 100$	95% CI
preserve	0.316	0.285–0.347	wicked	0.007	0.000–0.013
purity	0.280	0.247–0.313	taint	0.008	0.000–0.013
pristine	0.266	0.227–0.305	whore	0.010	0.004–0.016
immaculate	0.264	0.238–0.290	dirty	0.011	0.002–0.020
sacred	0.247	0.209–0.286	exploitations	0.012	0.006–0.018
	⋮			⋮	
virgins	0.149	0.111–0.188	ruins	0.030	0.022–0.039
abstinence	0.144	0.114–0.173	stainless	0.029	0.021–0.036
sterilization	0.075	0.043–0.107	lax	0.027	0.019–0.036
cleans	0.059	0.028–0.090	sickle	0.027	0.019–0.034
innocent	−0.051	−0.080–0.022	sinner	0.026	0.018–0.034

Note: Values are scaled by a factor of 100 for presentability. CI = confidence interval.

another set of 17 words for the DTM. We then add the same 17 nonanchor words as before. So, for the DTM, the first 17 words are the held-out sanctity words, the next 17 are the words unrelated to the sanctity ↔ degrade dimension, and the last 17 are the heldout degrade words. The rest of the procedure is equivalent with the compound concept example above.[10]

The right panels of Figure 3 show the results with the same visual parameters as the compound concept illustration in the left panels. The global relcos are again positive, as we would expect, and statistically distinguishable from zero at least at $\alpha = 0.001$ (right-tailed) for each of the three sets.

Table 3 lists the top and bottom five anchor word contributions for each pole of the direction relative to the orthogonal nonanchor words. *Innocent* has a value less than 0 and might thus be a candidate for removal from $\mathbf{\Psi}_{+,k}$. In fact, removing *innocent* increases the global reliability coefficient of these anchor lists from 0.290 to 0.291–a trivial, yet positive, increase, as we would expect.

Table 4. Final Expert Anchor Word Lists.

Moral Foundation	$\lvert\Psi_+\rvert$	Example Words	$\lvert\Psi_-\rvert$	Example Words
Care ↔ harm	59	safety, guard, shield, defence, guardian, protective	71	harmful, endanger, abandon, suffer, ruin, exploit
Sanctity ↔ degradation	35	purity, upright, decent, clean, refined, sacred	68	disgust, stains contagious, dirt, obscene, filth
Fairness ↔ cheating	28	fair, equal, equity, honest, justice, impartial	31	unfair, segregation, unjust, bias, exclude, discriminate
Loyalty ↔ betrayal	54	family, together, community, devotion, clique, loyalty	32	betrayal, treason, spy, betrayer, foreigner, enemy
Authority ↔ subversion	84	preserve, authority, status, leaders, ranking, obey	31	rebellion, protest, riot, rebels, oppose, illegal

Note: Words may appear across multiple anchor word lists.

Word-Level Validation

We now present a series of analyses validating relco for anchor words defining compound concepts, semantic centroids, and semantic directions. We first overview the materials for the condition manipulation—a series of expert- and crowdsourced moral foundation dictionaries—and the design and analytic plan before presenting the results for each of the three relation types. We focus on the case of orthogonal nonanchor words.

Hypothesis and Data

We validate relco by testing whether it can distinguish between anchor lists of varying quality. Our hypothesis is as follows:

> *Hypothesis 1:* Global relcos for higher confidence anchor word lists will be higher than global relcos for lower confidence anchor word lists.

By *confidence*, we mean the degree of consensus that words index the concept of interest. The idea is that if our proposed coefficient accurately measures the extent to which an anchor word list is reliable (because more respondents *agree* that it indexes the concept of interest), then it should be higher for lists we expect to be more reliable and lower for lists that are systematically manipulated to be less reliable. This setup requires validation anchor lists that we can subject to reliability manipulation.

We use two sources of moral foundations dictionaries. Our highest confidence anchor lists are the expert-curated Moral Foundations Dictionaries (Graham et al. 2009) derived from systematic content analysis of religious sermons by leading moral foundation theorists. These represent the gold standard for reliability. The dictionaries are stemmed for compatibility with LIWC software; we expanded these stems to full words using the procedures detailed in Appendix B. Table 4 shows the final expert anchor lists for each foundation's virtue and vice poles, with list sizes ranging from 28 to 84 words per pole.

For lower confidence conditions, we use the crowdsourced extended Moral Foundations Dictionary (Hopp et al. 2021), in which a large collection of English words received weights indicating the extent to which U.S.-based annotators collectively agreed each word indexed a given moral foundation. We created anchor lists at the 90th, 75th, 50th (median), and 25th

percentiles of these crowdsourced weights, maintaining list sizes equal to the expert-curated word list (see Table 4). This yielded five conditions per foundation: expert (highest confidence) down to the 25th percentile (lowest confidence). We used the pretrained fastText English word embeddings (Mikolov et al. 2018), which consist of 2 million word vectors trained on 600 billion tokens from the Common Crawl.

Design and Analytic Plan

We calculated the global relco for compound concepts, centroids, and directions across all five foundations and five confidence conditions (75 total global relcos each based on 1,000 simulation runs). To test whether relcos varied by confidence condition, we averaged concept engagement scores row-wise across runs, then used two-way analyses of variance (ANOVAs) to assess whether the relationship between concept engagement and anchor inclusion score (i.e., the global relco) varied significantly by condition. *Post hoc* Tukey honestly significant difference pairwise tests identified which coefficient pairs differed significantly. Appendix B provides full design specifications.

Results

All 15 two-way ANOVA interaction effects were statistically significant,[11] indicating that global relcos likely vary by anchor list confidence across all foundations and relation types. Figure 4 depicts these effects. The pattern is consistent: expert-curated lists (purple lines) show the steepest slopes, indicating highest reliability. As confidence decreases from expert to the 25th percentile, slopes systematically attenuate (purple all the way to yellow). The hypothesis therefore finds support. As the anchor list confidence decreases, the global relco also decreases. This is evidence that the proposed metric is capturing differences in anchor list reliability.

Baselines

We have established that relco captures anchor list reliability. However, unlike our validation test, most analysts with real-world empirical use cases will not have multiple anchor lists to adjudicate between. Instead, they will have one anchor list and therefore one relco. Thus, how does an analyst know a good relco when they see one—in absolute terms, not relative to competing relcos? And is zero a reasonable null hypothesis value?

We used simulations to establish some reasonable baselines for comparing relco effect sizes and for null hypothesis testing for each of the three relation types. For compound concepts and semantic centroids, we drew K vectors of real numbers from a normal parent distribution with a mean and variance equal to the mean (of vector-specific means) and variance of the fastText word embedding space. Half of those K (which we call k) were randomly drawn. The other half were drawn such that each vector was a perfect linear transformation of one another, meaning each pair had a Pearson's correlation of exactly 1. The random half of K were treated as the nonanchor words; the perfectly correlated half were treated as the anchor words. This simulated embedding space represented the ideal case where the anchor words indexing a latent relation are perfectly predictive of one another and orthogonal with the nonanchor words. We then repeated this process 100 times, each time introducing a small amount of random noise into the anchor word vectors. This resulted in a series of relcos that ranged from the best-case scenario, where the anchor words are perfectly reliable relative to random words, to the worst-case scenario, where the anchor words are as good as random (i.e., completely unreliable).

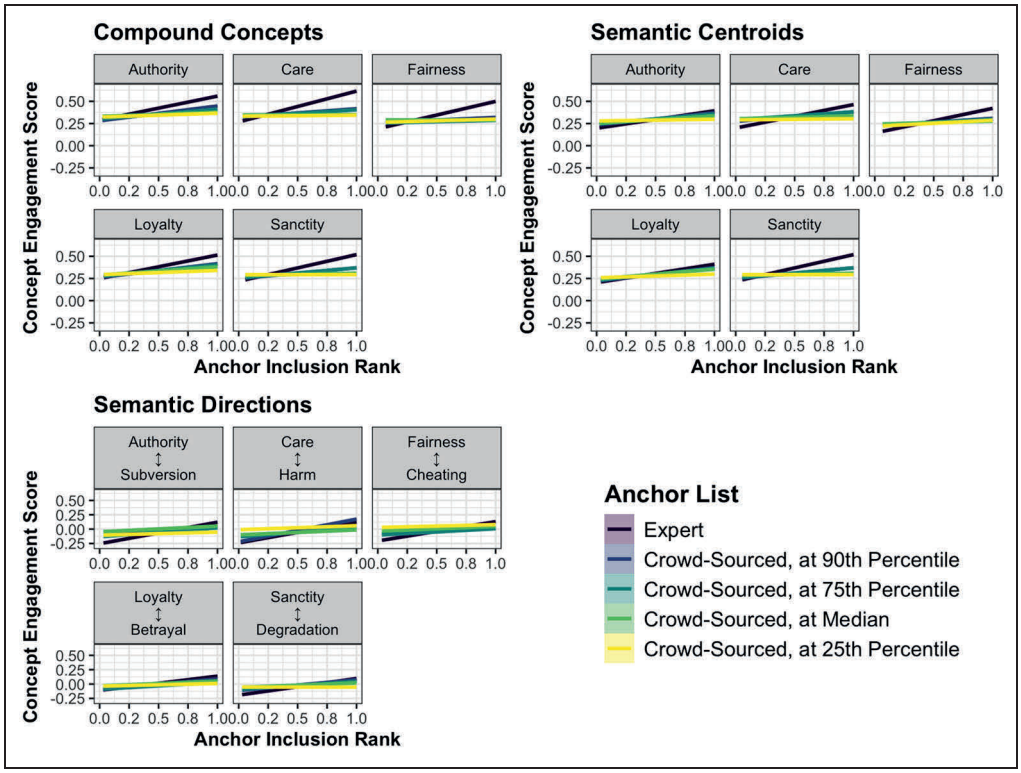

Figure 4. Word-level validation results.
Note: Each slope is derived after rank-wise averaging across 1,000 runs.

Each of these 100 scenarios was repeated a large number of times. The case was identical for the semantic direction simulations, but this time splitting K into three groups: one nonanchor set randomly drawn, one anchor word set drawn such that each pair of vectors was correlated at 1, and another anchor set drawn such that each pair was correlated at 1 and correlated with the other anchor set at -1. Finally, this entire procedure was repeated with anchor list lengths equal to 5, 10, 20, 30, 40, and 50, to ensure the results are stable at different anchor list sizes. More details on the simulation procedure are in Appendix C.

The relco distributions are visualized in Figure 5. The relation-specific distributions are virtually identical to one another in terms of starting values, ending values, and shape. This suggests relcos at different levels of anchor list reliability, from perfect to random, are not sensitive to the size of the anchor list. The range of values *do* vary by relation type, however, suggesting relation-specific baselines are needed.

Figure 6 reports the relation-specific median (of mean) relcos from Figure 5 by the quintiles of the average correlation between the anchor list words. The first bar shows the median relco in the simulated embedding spaces where each of the anchor word vectors are completely or nearly completely orthogonal to one another, and the last bar is the median relco in the simulated spaces where the anchor words are perfectly or nearly perfectly correlated with one another (and orthogonal to the nonanchor words). The relcos for the bottom three quantiles per relation are our suggested baselines, which we summarize more explicitly in Table 5.

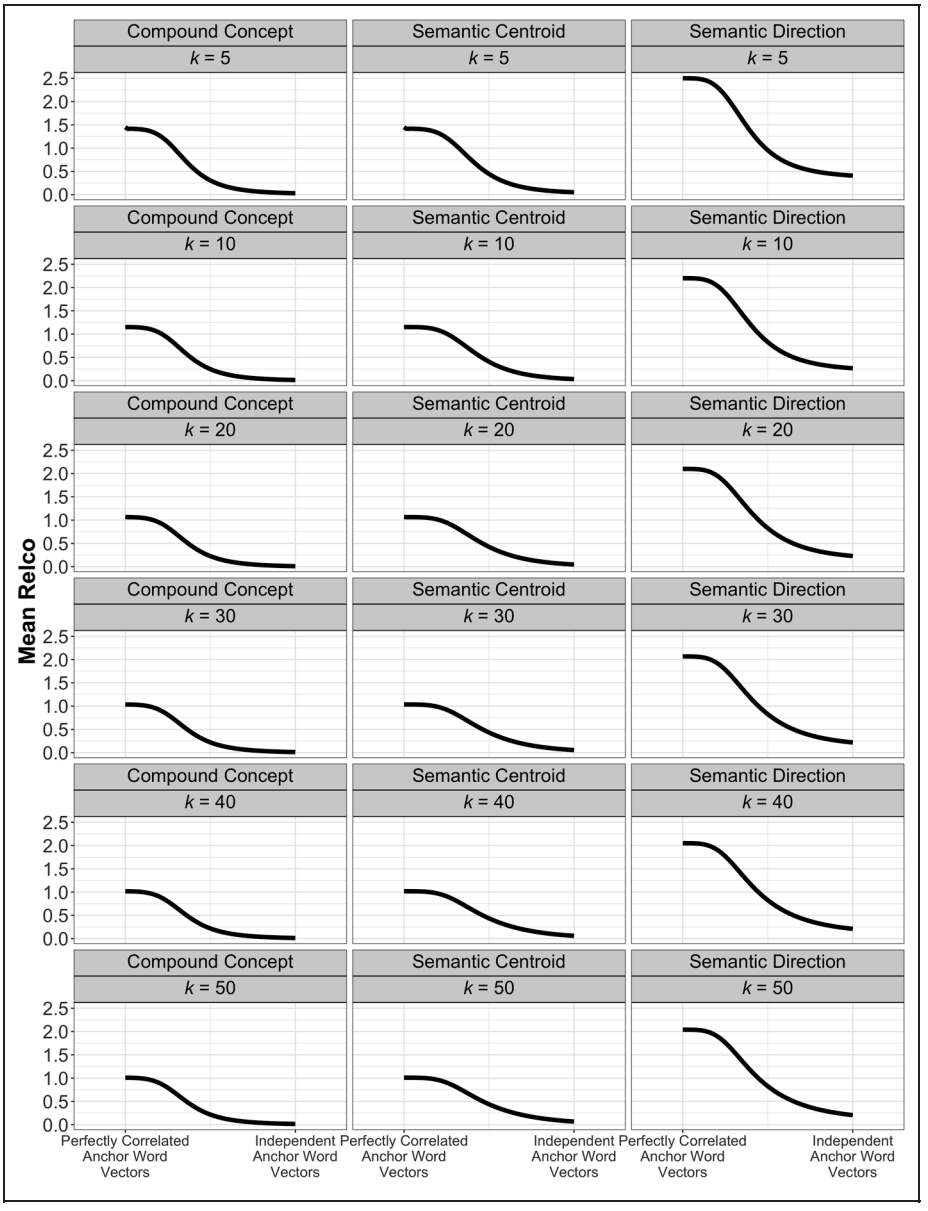

Figure 5. Distributions of simulated mean anchor reliability coefficient (relcos) at different levels of anchor list correlation.

Note: The *x*-axis indicates the correlation between the simulated anchor list vectors. The values range from 1 (where the anchor words are perfectly correlated, the nonanchor words are independent of one another, and the anchor words and nonanchor words are independent of one another) to 100 (where all anchor and nonanchor words are independent of one another). The relco simulations are run per *x*-axis value were each run is equal to the length of the anchor list set and then averaged, with each relco simulation based on 100 runs.

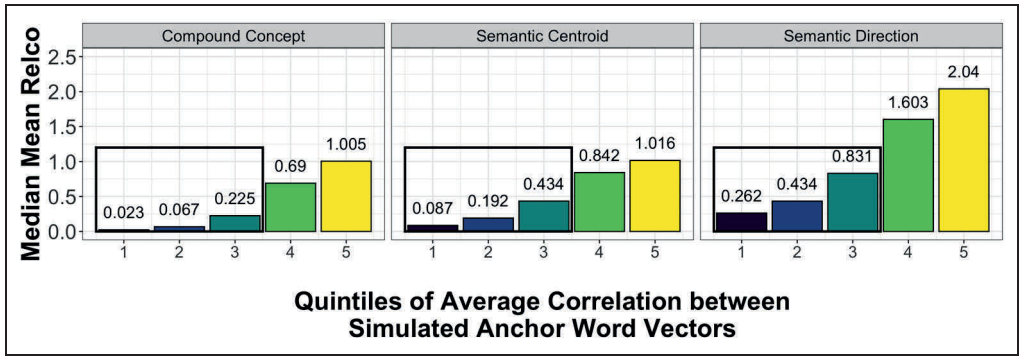

Figure 6. Suggested null hypothesis/effect size thresholds (in boxes).
Note: The quintile groups are organized such that groups 1 and 5 are the word vector spaces with anchor list correlations in the bottom 20 percent and top 20 percent of the overall distribution, respectively.

Table 5. Suggested Baselines for Effect Size Comparisons and Null Hypothesis Testing.

Relation Type	Reliability Level	Baseline
Compound concept	Weak reliability	0.03
	Moderate reliability	0.07
	Strong reliability	0.23
Semantic centroid	Weak reliability	0.09
	Moderate reliability	0.19
	Strong reliability	0.43
Semantic direction	Weak reliability	0.26
	Moderate reliability	0.43
	Strong reliability	0.83

Document-Level Validation

Beyond distinguishing between anchor lists of varying quality, we assess whether relco consistently informs anchor list construction for downstream classification tasks. We do this by varying the mean global relco for a series of anchor lists, using those lists to create semantic centroids, and classifying the documents in a real corpus on the basis of how close their words are to each of these centroids under various task specifications. The goal is to assess how *consistently* the relco of each list relates to the performance of the list as a classifier. If relco is consistently associated with better classification performance, then we can say relco is a robust indicator of anchor list quality across various task specifications that an analyst might use.

Hypothesis and Data

Our hypothesis for the following binary document classification task is as follows:

Hypothesis 2: Anchor sets with higher average global relcos will consistently result in classifications with higher agreement with document labels than will lower average global relcos, regardless of task specification.

Our corpus is a collection of 104,665 tweets from across 58 public health agencies in the United States, all posted between January 2020 and May 2023. These dates correspond to just before the World Health Organization declared coronavirus disease 2019 (COVID-19) a global pandemic, up to and including the month the World Health Organization declared the pandemic over.[12] Analyses were performed on a test data set, with all tweets manually coded as either 0 (not related to the COVID-19 pandemic) or 1 (related to the COVID-19 pandemic). More details on data collection, manual coding, and validation are in Appendix D.

Design and Analytic Plan

We began with a large 51-word dictionary of pandemic-related terms for COVID-19 classification. To test robustness across task specifications, we systematically varied four parameters. First, we used four embedding spaces. Three of the spaces were static: word2vec (CBOW) embeddings pretrained on the Google News corpus (Mikolov, Chen, et al. 2013), pretrained fastText embeddings, and GloVe embeddings (Pennington, Socher, and Manning 2014) locally trained on our 104,664-tweet corpus. The fourth was a transformer-based contextualized vector space: embeddings from BERTweet (Nguyen, Vu, and Nguyen 2020), a BERT model fine-tuned on 850 million tweets with COVID-19 overrepresentation. For each embedding space, we calculated each anchor list's global relco (1,000 runs, using randomly drawn nonanchor words from the Norvig Google Trillion Word Corpus; Norvig 2009).

Second, we varied the number of anchor words used to create the semantic centroids: $K = 3$, 4, 5, and 6. Third, we set $K = 4$ and trimmed the initial dictionary to $N = 51$, 41, 31, and 21 words. In both cases, we created semantic centroids for each set of anchor terms and measured each tweet's engagement to those centroids using CMD (standardized). Fourth, we then classified a tweet as pandemic related or not using three thresholds (0, 0.25, and 0.5). For each configuration of these four parameters, we measured classification agreement with manual labels using $F1$ scores, then regressed $F1$ on the average relco of the K words defining each classifier.

Critically, we assess whether higher average relcos *consistently* predict better performance, regardless of variations in these four parameters (K, N, classifier thresholds, or embedding space). We do not assess overall classifier performance. We therefore performed only minimal preprocessing and model adjudication. Full details on embedding specifications, dictionary content, and classification procedures are in Appendix D.

Results

The results are shown in Figures 7 and 8. In short, we find support for our hypothesis across all embeddings, anchor set sizes, classification thresholds, and different size starting dictionaries. If the average relco of the anchors used to define a semantic centroid is low, the classifier performs worse than if the average relco is higher. All but three slopes across the two figures are statistically distinguishable from zero (or less than zero) at least at $\alpha = 0.05$ using Bonferroni-corrected right-tailed t tests.[13]

Conclusions

Defining semantic relations from word vector spaces involves selecting a few "seed" or "anchor" words. Available tests of the reliability of anchor sets are specific to the case of relations formed from juxtaposing antonym pairs. We therefore propose a reliability metric that is easily interpretable and agnostic to the type of relation.

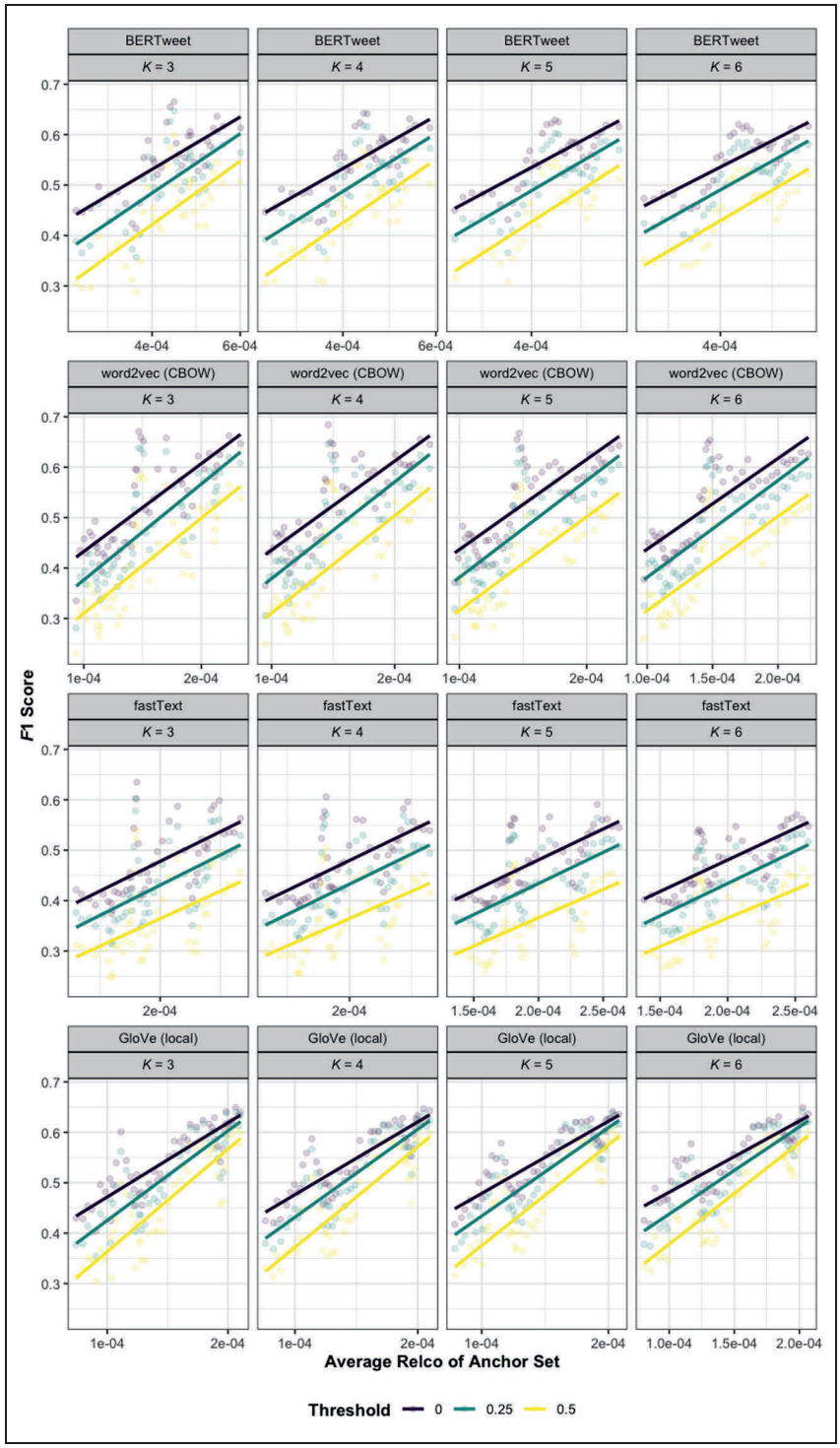

Figure 7. Average anchor reliability by *F*1 score, varying *K*.
Note: K refers to the length of the moving window of terms used to define a semantic centroid.

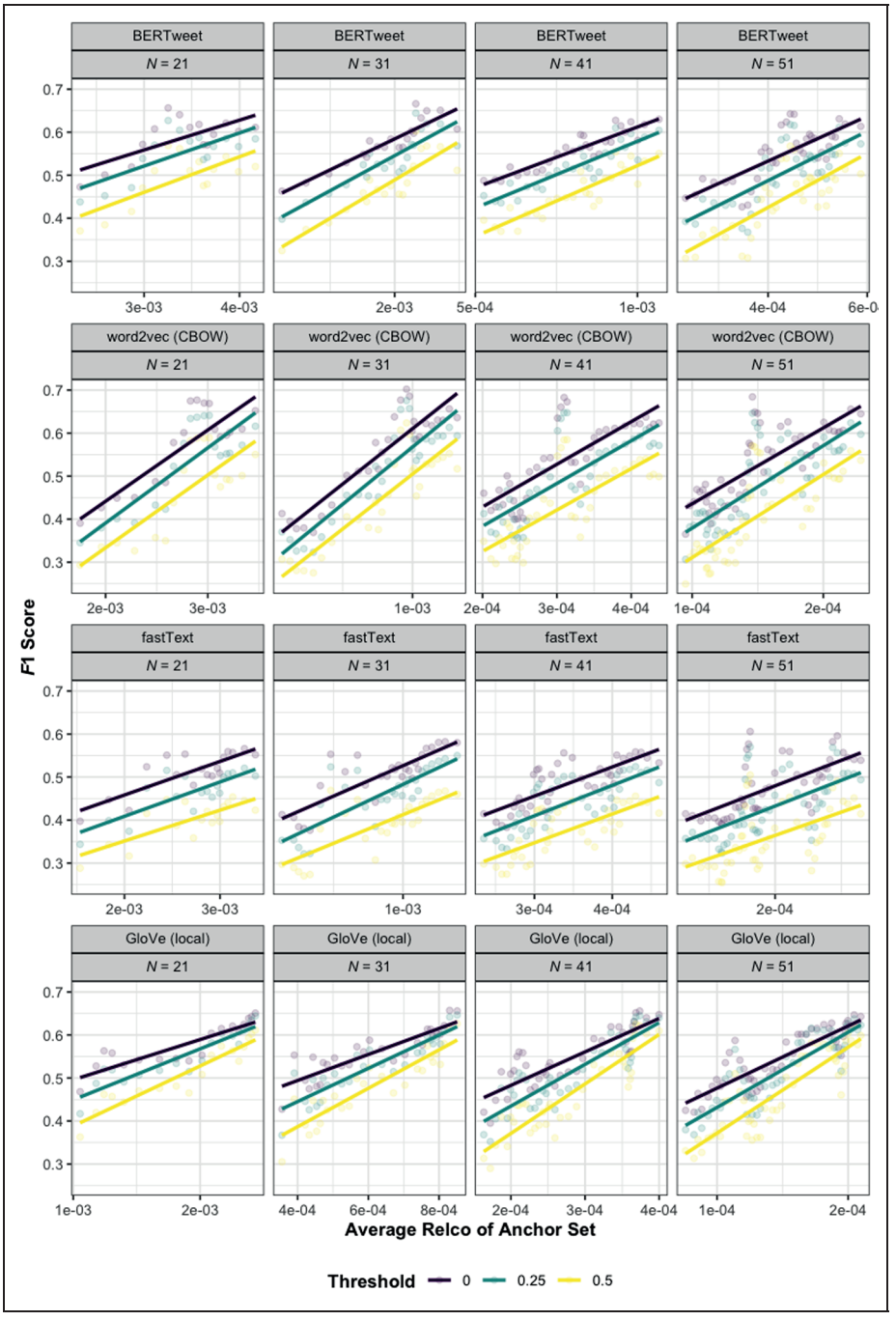

Figure 8. Average anchor reliability by *F*1 score, varying *N*.
Note: N refers to the size of the starting anchor list/dictionary.

The metric, which we call the anchor reliability coefficient, is found by creating simulated documents that sequentially shift more of their tokens from relation-relevant anchor terms to randomly drawn terms and then regressing the documents' similarity to an induced relation on the anchor inclusion score of the documents. We demonstrated the validity of this metric at the word level and document level. These tests suggest the relco is a useful metric, alongside others (e.g., Antoniak and Mimno 2021; Boutyline and Johnston 2025; Ethayarajh et al. 2019b), for guiding the construction of anchor sets when using word embedding models in cultural analysis.

Future Directions

One use case this article did not show was how relco could be used to find *better* anchor words. An analyst can use relco to construct more reliable anchor lists in an iterative fashion by (1) removing words recommended for potential exclusion by relco, (2) substituting in new words on the basis of theoretical knowledge or intuition, (3) calculating relco again, and (4) repeating this process until relco seems to plateau and with words that each meaningfully contribute to the reliability score.

This "*ad hoc*" approach may not be ideal, however. Work on keyword optimization for document retrieval suggests that humans are better at recognizing keywords than they are at coming up with them; that is, people are better at *remembering keywords* than they are at *recalling keywords* (King, Lam, and Roberts 2017:974). Any two people are unlikely to come up with the same keywords for a concept, signaling a serious issue with keyword and anchor word reliability when word lists are put together *post hoc* with human recall. Even worse, different word lists can have major downstream biases for analysis results (King et al. 2017:974). These findings suggest that an algorithmic approach to anchor word selection is likely a step in the right direction. One approach that leverages relco would include developing an algorithm to remove words from the user-supplied candidate anchor list with statistically nonsignificant contributions to relco, but then iterating relco over a large vocabulary (e.g., the Norvig Google Trillion Word Corpus [Norvig 2009] we used here) and retaining words that have a relco contribution equal to or above some threshold set by the analyst.

Future work could also compare different document aggregation techniques relative to the one used here (CMD). Such techniques summarize the extent any given combination of nonanchor words and one-half the anchor words engage the concept, as defined by the other half of the anchor words. For example, an alternative approach is to calculate the document centroid for each artificial document (Lix et al. 2022). It is less clear, though, how document centroids might need to be adapted to account for the bipolar structure of semantic directions. Importantly, however, any document-level representation derived from an embedding space can be dropped into the relco procedure.

Limitations

The size of the anchor list split was effectively held constant across all illustrations and validation tests as a 50:50 split: half the words to define the concept, and the other half for the artificial document. Although we believe that a 50:50 split is a reasonable default absent some *a priori* reason to favor a different split, we do acknowledge that such reasons may exist. We encourage relco users interested in alternative splits to consider how this might affect their baselines for effect size comparisons and statistical testing.

We end with a word of caution about using the metric—indeed any reliability metric. As stated before, relco should be treated as a statistical heuristic that supports, not replaces, expert qualitative assessments of the extent to which anchor list words lexicalize the same underlying semantic relation. One criticism of relco might be that analysts could "hack" it[14]—for example, perhaps picking anchor list words that are extremely narrow synonyms (but do not maximally cover the meaning of the intended relation) to artificially inflate the coefficient value. This is possible, but we stress this is not unlike "hacking" of other statistical metrics. *P*-hacking is the classic example, as well as accuracy in supervised learning tasks where skewed distributions across classes can result in a model that is only good at predicting particular classes. Relco should be used as a decision-making *guide*, not the final decision in and of itself.

Authors' Note

No artificial intelligence tools were used in the writing of this article.

Acknowledgments

Previous versions of this paper were presented at the Artificial Intelligence and Integrated Computer Systems Workshop at Linköping University, the International Roundtable on Computational Social Science at Linköping University, the Stockholm University Computational Sociology Working Group, and the 17th Annual Conference International Network of Analytical Sociology. We thank the editors and anonymous reviewers at *Sociological Methodology* for their invaluable comments and suggestions.

ORCID iDs

Marshall A. Taylor https://orcid.org/0000-0002-7440-0723
Dustin S. Stoltz https://orcid.org/0000-0002-4774-0765
Heather Harper https://orcid.org/0000-0002-0206-517X
Sumanth Reddy Nandhikonda https://orcid.org/0009-0002-5034-4904

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the National Science Foundation under award 2346727. Some of this research was conducted at the Swedish Excellence Center for Computational Social Science, which is funded by the Swedish Research Council (award 2022-06611).

Declaration of Conflicting Interests

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

Data Availability Statement

A replication repository for this paper is available at https://gitlab.com/mtaylor_soc/relco_repo.

Supplemental Material

Supplemental material for this article is available online.

Notes

1. For example, semantic coherence (Mimno et al. 2011), exclusivity (Airoldi and Bischof 2016), or term coherence (Sievert and Shirley 2014).
2. Boutyline and Johnston also consider a measure of synonymy as a potential reliability metric, as it captures the extent to which two anchor words for a given pole are identical in the embedding space. However, they find this measure (and a related measure of antonymy) adds little explanatory value after accounting for parallelism when predicting axis accuracy.
3. The relco, though, can also be calculated for the same semantic directions that parallelism assesses, as we will illustrate in later sections.
4. The suffixes of the dictionary items are normalized to work with the Linguistic Inquiry and Word Count (LIWC) software program (Pennebaker, Francis, and Booth 2001), so the list could be expanded to cover more variants of words. For example, *sacred** —as it is shown in the actual dictionary—could be expanded to include both *sacred* and *sacredness* in non-LIWC applications such as this one. We adopt a more systematic approach to expanding these suffixes in the "Word-Level Validation" section.
5. This, of course, is a question of validity, not reliability.
6. The possible range of CMD values varies as a function of a number of factors, especially the choice of either cosine distance or Euclidean distance as the travel distance metric (and, for Euclidean distance, whether the word embeddings are first normalized). We use the text2map package in the R statistical computing environment to calculate CMD scores (Stoltz and Taylor 2022), which in turn relies on the text2vec package (Selivanov, Bickel, and Wang 2023) to optimize the relaxed word mover's distance—the algorithmic foundation of CMD. The text2map implementation of relaxed word mover's distance uses cosine distance; thus CMD in text2map has a theoretical range of $[-1,1]$. In our experience, though, the observed range is usually closer to [0,1], likely for reasons similar to the anisotropic geometric properties of (static and contextualized) word embedding spaces that lead empirical cosine similarities to also stay within the [0,1] interval (Liang et al. 2021; Mu, Bhat, and Viswanath 2017).
7. The norm itself is equal to $k+1$ for compound concepts and semantic centroids and $k+k+1$ for semantic directions.
8. For semantic directions, we would rewrite this to say, accounting for every possible random partition (without replacement) of $\mathbf{\Psi}_{+x}$ and $\mathbf{\Psi}_{-x}$ and every possible permutation of $\mathbf{M}$ column word order across the $\mathbf{\Psi}_{+y}$, $\mathbf{\Psi}_{-x}$, and $\mathbf{\Upsilon}$ lists.
9. We have no experimental evidence to favor any particular penalty factor, although the value should be constrained to be between [0,1] and larger factors should result in each word being chosen at least once across fewer n runs.
10. The conceptually distinct nonanchor list was constructed by randomly sampling 17 words from both authority and subversion anchor lists.
11. All F ratios were significant at least at $\alpha = 0.001$.
12. The agencies include federal agencies (e.g., the Environmental Protection Agency and various Centers for Disease Control and Prevention accounts) and 44 of the 50 states (Wyoming, Rhode Island, West Virginia, New Jersey, Massachusetts, and New York are missing from the data).
13. Only the BERTweet slopes at $N = 21$ in Figure 8 have $p \geq .05$.
14. We thank a reviewer for pointing out this possibility.

References

Airoldi, Edoardo M., and Jonathan M. Bischof. 2016. "Improving and Evaluating Topic Models and Other Models of Text."*Journal of the American Statistical Association* 111(516):1381–403.

Antoniak, Maria, and David Mimno. 2021. "Bad Seeds: Evaluating Lexical Methods for Bias Measurement." Pp. 1889–1904 in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*

Processing, edited by C. Zong, F. Xia, W. Li, and R. Navigli. Kerrville, TX: Association for Computational Linguistics.

Arseniev-Koehler, Alina. 2022. "Theoretical Foundations and Limits of Word Embeddings: What Types of Meaning Can They Capture?" *Sociological Methods & Research* 53(4):1753–93.

Arseniev-Koehler, Alina, and Rachel Kahn Best. 2025. "Disease Frames and Their Consequences for Stigma and Medical Research Funds."*Social Science & Medicine* 372:117949.

Arseniev-Koehler, Alina, and Jacob G. Foster. 2022. "Machine Learning as a Model for Cultural Learning: Teaching an Algorithm What It Means to Be Fat."*Sociological Methods & Research* 51(4): 1484–1539.

Best, Rachel Kahn, and Alina Arseniev-Koehler. 2023. "The Stigma of Diseases: Unequal Burden, Uneven Decline."*American Sociological Review* 88(5):938–69.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." Pp. 4349–47 in *Advances in Neural Information Processing System, Vol. 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. New York: Curran Associates.

Boutyline, Andrei, and Alina Arseniev-Koehler. 2025. "Meaning in Hyperspace: Word Embeddings as Tools for Cultural Measurement."*Annual Review of Sociology* 51:89–107.

Boutyline, Andrei, and Ethan E. Johnston. 2025. "Forging Better Axes: Evaluating and Improving the Reliability of Semantic Dimensions in Word Embeddings." SocArXiv. Retrieved April 28, 2026. https://osf.io/preprints/socarxiv/576h3.

Breiger, Ronald L. 1974. "The Duality of Persons and Groups."*Social Forces* 53(2):181–90.

Carmines, Edward G., and Richard A. Zeller. 1979. *Reliability and Validity Assessment*. Beverly Hills, CA: Sage.

Daenekindt, Stijn, and Julian Schaap. 2022. "Using Word Embedding Models to Capture Changing Media Discourses: A Study on the Role of Legitimacy, Gender and Genre in 24,000 Music Reviews, 1999–2021."*Journal of Computational Social Science* 5(2):1615–36.

Durrheim, Kevin, Maria Schuld, Martin Mafunda, and Sindisiwe Mazibuko. 2022. "Using Word Embeddings to Investigate Cultural Biases."*British Journal of Social Psychology* 62(1):617–29.

Ethayarajh, Kawin, David Duvenaud, and Graeme Hirst. 2019a. "Towards Understanding Linear Word Analogies." Pp. 3253–62 in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by A. Korhonen, D. Traum, and L. Màrquez. Kerrville, TX: Association for Computational Linguistics.

Ethayarajh, Kawin, David Duvenaud, and Graeme Hirst. 2019b. "Understanding Undesirable Word Embedding Associations." Pp. 1696–1705 in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by A. Korhonen, D. Traum, and L. Màrquez. Kerrville, TX: Association for Computational Linguistics.

Graham, Jesse, Jonathan Haidt, and Brian A. Nosek. 2009. "Liberals and Conservatives Rely on Different Sets of Moral Foundations."*Journal of Personality and Social Psychology* 96(5):1029–46.

Haidt, Jonathan, and Jesse Graham. 2007. "When Morality Opposes Justice: Conservatives Have Moral Intuitions That Liberals May Not Recognize."*Social Justice Research* 20(1):98–116.

Hopp, Frederic R., Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. "The Extended Moral Foundations Dictionary (eMFD): Development and Applications of a Crowd-Sourced Approach to Extracting Moral Intuitions from Text."*Behavior Research Methods* 53:232–46.

Johnson, Amy L. 2024. "Psychotic White Men and Bipolar Black Women? Racialized and Gendered Implications of Mental Health Terminology."*Social Science & Medicine* 352:117015.

Jones, Jason J., Mohammad Ruhul Amin, Jessica Kim, and Steven Skiena. 2020. "Stereotypical Gender Associations in Language Have Decreased over Time."*Sociological Science* 7:1–35.

Joseph, Kenneth, and Jonathan Morgan. 2020. "When Do Word Embeddings Accurately Reflect Surveys on Our Beliefs about People?" Pp. 4392–415 in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault. Kerrville, TX: Association for Computational Linguistics.

King, Garry, Patrick Lam, and Margaret E. Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text."*American Journal of Political Science* 61(4):971–88.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings."*American Sociological Review* 84(5):905–49.

Liang, Yuxin, Rui Cao, Jie Zheng, Jie Ren, and Ling Gao. 2021. "Learning to Remove: Towards Isotropic Pre-trained BERT Embedding." Pp. 448–59 in *Artificial Neural Networks and Machine Learning—ICANN 2021*, edited by I. Farkaš, P. Masulli, S. Otte, and S. Wermter. Cham, Switzerland: Springer.

Lix, Katharina, Amir Goldberg, Sameer B. Srivastava, and Melissa A. Valentine. 2022. "Aligning Differences: Discursive Diversity and Team Performance."*Management Science* 68(11):8430–48.

McCumber, Andrew, and Adam Davis. 2024. "Elite Environmental Aesthetics: Placing Nature in a Changing Climate."*American Journal of Cultural Sociology* 12(1):53–84.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." Retrieved April 28, 2026. https://arxiv.org/abs/1301.3781.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. "Advances in Pre-training Distributed Word Representations." Pp. 52–55 in *Proceedings of the International Conference on Language Resources and Evaluation*, edited by N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, et al. Paris: European Language Resources Association.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. "Linguistic Regularities in Continuous Space Word Representations." Pp. 746–51 in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by L. Vanderwende, H. Daumé III, and K. Kirchoff. Kerrville, TX: Association for Computational Linguistics.

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing Semantic Coherence in Topic Models." Pp. 262–72 in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, edited by R. Barzilay and M. Johnson. Kerrville, TX: Association for Computational Linguistics.

Mohr, John W., and Vincent Duquenne. 1997. "The Duality of Culture and Practice: Poverty Relief in New York City, 1888–1917."*Theory and Society* 26(2/3):305–356.

Mu, Jiaqi, Suma Bhat, and Pramod Viswanath. 2017. "All-but-the-Top: Simple and Effective Postprocessing for Word Representations." arXiv. Retrieved April 28, 2026. https://arxiv.org/abs/1702.01417.

Nelson, Laura K. 2021. "Leveraging the Alignment between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century U.S. South."*Poetics* 88:101539.

Nguyen, Dat Quoc, Thanh Vu, and Anh Tuan Nguyen. 2020. "BERTweet: A Pre-trained Language Model for English Tweets." Pp. 9–14 in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, edited by Q. Liu and D. Schlangen. Kerrville, TX: Association for Computational Linguistics.

Norvig, Peter. 2009. "Natural Language Corpus Data." Pp. 219–42 in *Beautiful Data: The Stories Behind Elegant Data Solutions*, edited by T. Segaran and J. Hammerbacher. Sebastopol, CA: O'Reilly Media.

Pennebaker, James W., Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count: LIWC 2001*. Mahwah, NJ: Lawrence Erlbaum.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." Pp. 1532–43 in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by A. Moschitti, B. Pang, and W. Daelemans. Kerrville, TX: Association for Computational Linguistics.

Pouliot, Vincent, and Scott Rovert Patterson. 2024. "Domesticating Wealth Inequality."*Global Studies Quarterly* 4(2):ksae023.

Rodriguez, Pedro L., and Arthur Spirling. 2022. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research."*Journal of Politics* 84(1):101–115.

Schoon, Eric W., David Melamed, and Ronald L. Breiger. 2024. *Regression Inside Out*. Cambridge, UK: Cambridge University Press.
Selivanov, Dmitry, Manuel Bickel, and Qing Wang. 2023. "text2vec: Modern Text Mining Framework for R." R Package Version 0.6.4. Retrieved April 28, 2026. https://CRAN.R-project.org/package=text2vec.
Sievert, Carson, and Kenneth E. Shirley. 2014. "LDAvis: A Method for Visualizing and Interpreting Topics." Pp. 63–70 in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, edited by J. Chuang, S. Green, M. Hearst, J. Heer, and P. Koehn. Kerrville, TX: Association for Computational Linguistics.
Stoltz, Dustin S., Marissa A. Combs, and Marshall A. Taylor. 2023. "Corpus Modeling and the Geometries of Text: Meaning Spaces as Metaphor and Method." Pp. 59–78 in *The Oxford Handbook of the Sociology of Machine Learning*, edited by C. Borch and J. P. Pardo-Guerra. Oxford, UK: Oxford University Press.
Stoltz, Dustin S., and Marshall A. Taylor. 2019. "Concept Mover's Distance: Measuring Concept Engagement via Word Embeddings in Texts."*Journal of Computational Social Science* 2(2):293–313.
Stoltz, Dustin S., and Marshall A. Taylor. 2021. "Cultural Cartography with Word Embeddings."*Poetics* 88:101567.
Stoltz, Dustin S., and Marshall A. Taylor. 2022. "text2map: R Tools for Text Matrices."*Journal of Open Source Software* 7(72):3741.
Stoltz, Dustin S., and Marshall A. Taylor. 2024. *Mapping Texts: Computational Text Analysis for the Social Sciences*. Oxford, UK: Oxford University Press.
Stoltz, Dustin S., Marshall A. Taylor, and Jennifer S. K. Dudley. 2024. "A Tool Kit for Relation Induction in Text Analysis."*Sociological Methods & Research* 54(2):565–604.
Taylor, Marshall A., and Dustin S. Stoltz. 2020a. "Concept Class Analysis: A Method for Identifying Cultural Schemas in Texts."*Sociological Science* 7:544–69.
Taylor, Marshall A., and Dustin S. Stoltz. 2020b. "Integrating Semantic Directions with Concept Mover's Distance to Measure Binary Concept Engagement."*Journal of Computational Social Science* 4: 231–42.
Taylor, Marshall A., and Dustin S. Stoltz. 2025. "A Workflow for Analyzing Cultural Schemas in Texts."*Journal of Mathematical Sociology* 49(1):1–24.
van Loon, Austin, and Jeremy Freese. 2023. "Word Embeddings Reveal How Fundamental Sentiments Structure Natural Language."*American Behavioral Scientist* 67(2):175–200.
Vann, Burrel, Jr. 2023. "The Framing of Marijuana in Black Newspapers."*International Journal of Press/ Politics* 30(1):370–97.
Voyer, Andrea, Zachary D. Kline, Madison Danton, and Tatiana Volkova. 2022. "From Strange to Normal: Computational Approaches to Examining Immigrant Incorporation through Shifts in the Mainstream."*Sociological Methods & Research* 51(4):1540–79.
Yoon, Hesu, and Andrew McCumber. 2024. "A Symbolic Hierarchy of Places: Global Inequalities in Tourism Narratives of the *New York Times* Travel Section."*Poetics* 102:101848.

Author Biographies

Marshall A. Taylor is an associate professor of sociology at New Mexico State University. He studies cultural knowledge: when, where, and why it is stable or changes, how it is structured, and how to best measure cultural knowledge in natural language and survey data using computational methods. He is the author, with Dustin S. Stoltz, of *Mapping Texts: Computational Text Analysis for the Social Sciences* (Oxford University Press 2024). His work has been published in peer-reviewed outlets such as *Sociological Methods & Research, Poetics, Sociological Theory, Political Behavior*, the *Journal of Mathematical Sociology*, and *Sociological Forum*.

Dustin S. Stoltz is an assistant professor of sociology and cognitive science at Lehigh University. His research explores how social structure, culture, and cognition shape ideas and evaluations. He is the author, with Marshall A. Taylor, of *Mapping Texts: Computational Text Analysis for the Social Sciences*

(Oxford University Press 2024). His work has been published in peer-reviewed outlets such as *Sociological Theory, Poetics*, the *Journal for the Theory of Social Behaviour*, and the *Journal of Computational Social Science*.

Heather Harper is an assistant professor of sociology at New Mexico State University. She specializes in public policy, political sociology, democracy studies, and computational text analysis. Much of her work involves analyzing policy design processes and outcomes as well as over time trends in policy content development. Current projects, funded by the National Science Foundation and the U.S. Department of Energy, focus on public health message contradictions, the spread of misinformation, and barriers to public support for climate technology investments.

Sanuj Kumar is a PhD student in computer science at New Mexico State University. His research centers on focused analysis of data, developing methods to detect subtle, hidden structures in large-scale, noisy text and to surface insights through the aspects that matter most to users. His interests span text mining, topic modeling, probabilistic models, variational autoencoders, diffusion models, and language models, with an emphasis on building interpretable representations that preserve aspect-level signals of interest. He is currently working on unsupervised abstractive multidocument summarization with language models, exploring ways to guide generation using aspect-aware, structure-seeking representations. Prior to this, his work included probabilistic and focused-analysis approaches for uncovering latent patterns and supporting user-driven exploration of complex data sets.

Sumanth Reddy Nandhikonda is a PhD student in computer science at New Mexico State University. His research interests are in machine learning, large language models, and applied data analysis. He is currently working on research projects involving educational research and artificial intelligence–driven methods.

Luke Burks is a PhD student in health equity sciences at New Mexico State University. His research interests include evaluating cognitive and social contributors of discrimination in disadvantaged populations, as well as exploring the relationship between perceptions of media messaging consistency and various predictive factors, evaluated using his qualitative and quantitative methodological expertise in public health. His research has been supported by the National Science Foundation and New Mexico State University.