AERA Open
July-September 2020, Vol. 6, No. 3, pp. 1–27
DOI: 10.1177/2332858420940312
Article reuse guidelines: sagepub.com/journals-permissions

Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks

Li Lucy
Dorottya Demszky
Patricia Bromley
Dan Jurafsky
Stanford University

Cutting-edge data science techniques can shed new light on fundamental questions in educational research. We apply techniques from natural language processing (lexicons, word embeddings, topic models) to 15 U.S. history textbooks widely used in Texas between 2015 and 2017, studying their depiction of historically marginalized groups. We find that Latinx people are rarely discussed, and the most common famous figures are nearly all White men. Lexicon-based approaches show that Black people are described as performing actions associated with low agency and power. Word embeddings reveal that women tend to be discussed in the contexts of work and the home. Topic modeling highlights the higher prominence of political topics compared with social ones. We also find that more conservative counties tend to purchase textbooks with less representation of women and Black people. Building on a rich tradition of textbook analysis, we release our computational toolkit to support new research directions.

Keywords: *artificial intelligence, case studies, content analysis, curriculum, data science, gender studies, history, natural language processing, race, textbooks, textual analysis*

Introduction

Recent methodological developments in a subfield of artificial intelligence—natural language processing (NLP)—offer great promise for shedding light on key questions about the social and political aspects of education. In particular, textbooks—conceptualized as artifacts of a dynamic cultural system—have long been a rich source of insight into schooling (FitzGerald, 1980; Loewen, 2008). We draw on a sample of 15 of the most widely used high school history textbooks in Texas, highlighting the insights that can be gained through the use of NLP or text data science methods. There are several potential contributions of NLP methods to curricular research. First, the methods allow us to measure complex concepts using larger sample sizes, which can shed new light on the scope and scale of trends in educational discourse. Second, there is greater capacity to analyze linguistic connections between words in the texts, which promotes attention to relational forms of meaning, allowing the discovery of topics and associations between concepts. Third, there is increased capability to systematically capture the way in which certain words are used to promote particular perspectives and frames. These measures, combined with the ability to use larger samples, allow researchers to analyze relationships between discourse and external factors in previously impractical ways.

While we see much promise in these approaches, it is important to understand computational textual analysis as a complement to, not a replacement for, more holistic analyses (e.g., ethnographic and case studies; Grimmer & Stewart, 2013; Nguyen, 2017). Indeed, we believe the flexibility of NLP tools can put greater responsibility on researchers to clearly specify the conceptual goals of research.

Our goal is to demonstrate methods for quantifying the content of textbooks that connect to the social scientific, policy, and practical aims of educational research. We do not provide a normative evaluation of textbooks, or a detailed analysis of why textbooks contain some kinds of content and not others. Instead, we use U.S. history textbooks from Texas as a case study to illustrate how NLP methods can answer research questions about depictions of historically marginalized groups that have been previously studied by textbook researchers using traditional methods. Our methodological and descriptive focus generates multiple avenues of future research that would hopefully increase interest in

developing better computational tools for this domain. These methods also support social scientific explanations of content and evidence-based policy prescriptions, which we reflect on in our conclusion. We release our toolkit for computational analyses of textbook content to the research community at https://github.com/ddemszky/textbook-analysis.

Textbook Research

Textbooks are central in educational research because they represent the "intended curriculum," sitting at the intersection between individual students and the macro forces of society, culture, and politics (Apple, 1992). Textbooks are also among the most widely used instructional technology around the world (Torney-Purta et al., 2001), and their availability and use positively influence student achievement (Fredriksen & Brar, 2015; Read & Bontoux, 2016). But, as is well known, textbooks are not neutral: their content is contested and reflects the power asymmetries and taken-for-granted beliefs of the underpinning culture (Moreau, 2010). These textbooks convey legitimated social and cultural values to students and impact students' perspectives of people and ethnicities different from themselves (Cornbleth, 2002; Greaney, 2006). Building on ongoing work in this area, we present methods that aid the study of depictions of gender, race, and ethnicity in contemporary history textbooks.

Educational researchers' understanding of the sources of textbook content and the mechanisms through which various discourses appear and spread have been limited by our methods. A main limitation of traditional methods is scalability. Most textbook content analyses continue to rely on a single researcher reading and hand coding textbooks (see Nicholls, 2003, or Pingel, 2010, for an overview of methods for textbook research), which is an extremely resource intensive endeavor to conduct at scale. For example, in a recent article in *American Journal of Sociology*, Morning (2008) hand codes 80 biology textbooks by manually searching for relevant segments of the book using index keywords. In some cases, scholars have multiple coders read books and conduct interrater reliability checks and test constructed measures for some level of statistical validity (e.g., Lerch et al., 2017). The accessibility of these constructed measures poses another limitation to hand-coding, since the annotation of subtle linguistic cues (e.g., the agency associated with certain verbs) requires training hand-coders to understand linguistic frameworks. Therefore, large textbook coding efforts may reduce their tasks to counting or identifying simple indicators. For example, Bromley et al. (2011) code a cross-national sample of over 500 social science textbooks for the presence or absence of discussions of "the environment". Due to the limitations of human coding a large, longitudinal, cross-national sample, the authors were unable to develop more nuanced indicators of environmental education.

Computational Approaches

NLP methods are popular in computational social science (see also Nguyen, 2017; O'Connor et al., 2011), and they have yielded important insights on textual data in the field of education. For example, they have been used to analyze online and in-person class discussions (Fesler et al., 2019; Lugini et al., 2018), topics in dissertation abstracts (Munoz-Najar Galvez et al., 2019), and disciplinary differences in students' academic writing (Crossley et al., 2017). A variety of NLP tools, such as Coh-Metrix (Graesser et al., 2014; McNamara et al., 2014), the Tool for the Automatic Analysis of Text Cohesion (TAACO; Crossley et al., 2016), and ReaderBench (Dascalu et al., 2014), have been used to characterize text cohesion, difficulty, and complexity in learning analytics and education data mining (Crossley & Kyle, 2018). These tools can enable educators to select education material suitable for students (Graesser et al., 2011) or analyze dialogue in digital learning environments at scale (Dowell et al., 2016). Other cases of NLP tools applied to educational texts include LightSIDE for automated essay evaluation (Mayfield & Rosé, 2013), TAALES for predicting lexical proficiency and word choice (Kyle et al., 2018), and Group Communication Analysis for detecting discussion participant roles (Dowell et al., 2019).

However, there has been less work on applying NLP to answer sociological questions in education.[1] Some early efforts apply machine counting of words to scanned textbooks, such as Lachmann and Mitchell (2014)'s study on depictions of war. A number of recent studies outside education have used NLP methods to study the reflection of gender and other social variables in text: Fast et al. (2016) look at gender stereotypes in online fiction; Hoyle et al. (2019) measured the association of adjectives and verbs with different genders in a million digitized books; Garg et al. (2018) quantified a century of gender and ethnic stereotypes using word representations learned from books, newspapers, and other texts; and Ash et al. (2020) examine the role of gender slant in judicial behavior using text written by judges. We build on this line of work examining depictions of social groups in texts (see also Field et al., 2019; Joseph et al., 2017; Ornaghi et al., 2019), extending NLP methods to textbooks.

Though NLP can achieve near-human performance on some linguistic tasks (Wang et al., 2019), its methods are still error-prone and subject to bias. So, its use in a field with high social impact, such as education, necessitates care when drawing conclusions (see Hovy & Spruit, 2016, or Olteanu et al., 2019 for an overview). One way to be careful is to strive for transparency and explainability when choosing methods. For example, lexicon-based approaches offer some interpretability by explicitly showing which words are counted. An overarching limitation of our work is that many machine learning models or resources are initially trained, or developed, on data from noneducational

genres such as news articles, and task performance may not transfer well to a different domain. Though it would be ideal to tailor these models to history textbooks in the future, these efforts will require extensive annotation of training data. In our methods section, we further elaborate on more specific limitations of NLP.

Overall, computational approaches are not a competitor to or replacement for traditional methods when tackling complex social phenomena. We seek to unite them with shared research goals and use the strengths of one method to assist potential weaknesses of another. In particular, NLP can only describe content, rather than prescribe it, so educators, ethicists, and social scientists should interpret results and determine if they align with the intended curriculum and proper goals of schooling.

Our Contributions

Our first contribution is that we provide quantitative and scalable measurements of textbook content. These methods, when applied across books, provide a more complete picture of the discourse around historical events and people in U.S. history education. The patterned exclusion of some views from history textbooks is well documented, but our analysis sheds new light on the scope and scale of this exclusion. For example, we find Latinx people are virtually absent from discussions of racial and ethnic groups in history textbooks in Texas, and nearly all famous figures discussed are White men in politics.

Our second key contribution is that by employing NLP methods that discover patterns in the co-occurrence of terms, we enable a relational approach to meaning relevant for textbook research. These methods uncover latent structures and networks of terms and can create a rich picture of how textbooks reflect social meaning. Moreover, they can help answer questions about the substantive nature of discourse; that is, what meanings are linked to certain terms or concepts? We show evidence that despite a move toward pedagogical approaches that focus on multiple perspectives of the past, history textbooks in Texas remain dominated by topics of formal politics. Our analyses also show that Black people are discussed using terms with lower levels of agency and power than other groups—a finding that also highlights the importance of combining substantive expertise with computational methods.

Finally, we demonstrate that a quantitative analysis of larger samples allows researchers to link patterns in the text to external social, political and cultural influences on and to consequences of education, at a scale that may be less feasible through broad hand-coding of fewer textbooks. Illustratively, we link textbook content to district purchasing patterns and districts' political leaning, which is one of many possible factors involved in the process of textbook creation and distribution. We find that although differences between textbooks are small compared with their similarities, districts in more Democratic counties tend to purchase textbooks that contain higher levels of representation of historically marginalized groups.

These contributions are motivated by our research questions, which we summarize in Table 1, along with the methods and resources used to answer each question. After outlining our data in the next section, we turn to describing each method in more detail. We then discuss the results of each question. We conclude by reflecting on the contributions and limitations of NLP methods for social science research in education.

Data

Texas Textbooks

We focus on textbooks used in Texas, which makes its district-level textbook purchase data available online in a unified format (Texas Education Agency, n.d.). Given that it has the second largest student population in the United States, with 5.4 million students enrolled in its K–12 public school system in 2017, Texas is a major textbook market for publishers, and so the state has a significant influence on U.S. textbook content. At the same time, the Texas Board of Education has been at the center of several textbook controversies. For example, the 2015 statewide social studies textbook adoption, driven by conservative ideology, triggered controversy over possible biases within curriculum content (Goldstein, 2020; Hutchins, 2011; Rockmore, 2015). Our dataset includes U.S. history textbooks widely purchased in Texas between 2015 and 2017. We select titles that occur in at least 10 district-level transactions. The final list of fifteen textbooks, including six combined volumes, is available in Textbook Sources. Additional details are available in Appendix C. Seven volumes were PDF files, and we extracted text directly from these files. As for the other volumes, we scanned and digitized them using ABBYY FineReader, which employs optical character recognition (OCR). We perform minimal post-processing on the text (Appendix D). Our textbook data contains a total of 7.6 million tokens, defined as strings of continuous characters between spaces or punctuation marks.

Demographic Data

We use geographic and student demographic data from the 2016–2017 school year collected by the National Center for Education Statistics (n.d.) Common Core of Data for public school districts to obtain textbook distribution data. In addition, to estimate the political leaning of each county, we use the two party vote shares from the 2016 elections, broken down by county (The New York Times, 2017).[2] In our analyses, we use estimates of Democratic vote-shares as an illustration of the types of external associations that

TABLE 1
Primary Contributions, Research Questions, Subproblems, Methods, and Resources

Research question(s)	Subproblem	Relevant method or resource
1. How much space is allocated to different groups?	Identifying people-related terms	WordNet (Miller, 1995)
	Identifying famous people	Named Entity Recognition, Wikidata
	Measuring space	Coreference resolution
2. How are different groups described?	Identifying descriptor words	Dependency parsing
	Comparing descriptors of different groups	Log odds ratio
	Measuring connotations of descriptors	National Research Council Lexicon (Mohammad, 2018); Connotation Frames (Rashkin et al., 2016; Sap et al., 2017)
	Comparing the association of words with different groups	Word embeddings
3. What are prominent topics and how are they related to groups of people?	Identifying topics	Topic modeling (Latent Dirichlet Allocation)
	Comparing the prominence of topics across books	Ratio of average topic probabilities

become possible with our methods. Future research designs could explore other mechanisms that might shape textbook distribution in a district, such as the demographics of a school or a school board.[3]

Method

Our goal is to apply NLP methods to examine depictions of historically marginalized groups in textbooks. We illustrate the methods we found most relevant, following the order of questions listed in Table 1.

Research Question 1: How Much Space Is Allocated to Different Groups?

Our methods in this section quantify the amount of textual space that different people and groups cover, in the spirit of the frameworks used in studies of multicultural curriculum to categorize textbook diversity. Previous traditional approaches in analyzing social studies textbooks have examined the presence and discussion of everyday, generic (nonnamed) people (e.g., *settler* or *farmer*) as well as named, famous individuals (e.g., Lincoln or Washington; Gordy & Pritchard, 1995; Schmidt, 2012). Researchers have measured diversity in textbooks by considering how much a minority group is mentioned relative to a majority, by examining whether texts portray minorities' roles as secondary or contributory, and by determining whether famous named people from a minority demographic are included (Banks, 2001; Gordy & Pritchard, 1995; Tetreault, 1986).

Identifying People-Related Terms. We identify common nouns that designate nonnamed people, such as *pioneer* or *Mexicans*, via WordNet, an English lexical database that encodes the meanings of words and relations between them (Miller, 1995), similarly to a thesaurus. We use the database to extract all hyponyms (subcategories) of *human, person, people*, and *social group*. For example, *pioneer* is a member of the hyponym chain *person* > *creator* > *originator* > *pioneer* in WordNet, and hence we obtain this term when we search for all hyponyms of *person*.

To evaluate how well this WordNet-based method performs, we also perform manual labeling. We use the spaCy package (Honnibal & Montani, 2017) to extract the heads of all noun phrases in the text that occur at least 10 times in all of our data.[4] This process yields around 12,000 unique noun heads. We manually combed through this list of heads to extract those common nouns that refer to people, resulting in 2,111 total terms. We find that our automatic WordNet-based method captures more than 95% of all manually identified nouns referring to people, and it captures 98 of the 100 most frequent nouns in our data referring to people—the exceptions being *group* and *majority*, which, in WordNet, are not hyponyms of people-referring terms, because they can refer to other entities as well. However, because in history textbooks we expect these two terms to refer to people, we still include them in our list for analyses, along with the remaining 5% of manually identified people-related terms not identified by the WordNet-based method.

Our list of people-related terms consists of 1,665 unmarked terms such as *engineer* or *family* as well as 446 terms specifying a demographic, including singular and plural forms of nouns (Appendix Table A1). To compare how different demographic groups are described, we manually categorize this list based on gender and ethnicity. Some of the nouns also have an adjectival sense (e.g., *Navajo community*), and therefore, when looking at specific mention of a people-term in text, we also count whether its adjectival markers are associated with a particular demographic. We also consider cases of intersectionality, such as *Black women*, which would be categorized as both *woman* and *Black*. For gender-based analyses we focused on women and men, because our dataset does not have many instances of other gender identities, and only three mentions of *transgender*.

Identifying Famous People. A textbook's discussion of social groups also involves mentioning individuals by name. Though the inclusion of a few standout individuals alone is not enough to label a textbook as diverse, their absence is a key sign that a textbook is missing crucial parts of American history (Banks, 2001). To identify named individuals, we use spaCy's named entity recognition (NER) tagger. A named entity is a proper noun describing a person, location, or organization, and taggers label these automatically. A manual evaluation of this NER tagger on our textbooks yielded an F1 score of .735 (Appendix D). The errors of NER and its potential biases when encountering names of different genders and backgrounds is an active area of research in NLP (Mehrabi et al., 2019), but our results indicate that, at a minimum, spaCy's pretrained tagger is accurate enough to motivate future work on adapting this model for textbook language.

To ensure that we do not double count individuals due to aliases (e.g., Franklin D. Roosevelt and Franklin Roosevelt), we pull aliases from the free knowledge base Wikidata and standardize these variations with their official Wikidata name. One limitation is that a knowledge base such as Wikidata, like its encyclopedic sister project Wikipedia, may contain less coverage of underrepresented individuals (Wagner et al., 2015). In addition, as NER tagging is somewhat noisy and captures a long tail of phrases that are not people, we only keep the top 100 most common NER-detected names, which restricts our focus to the people textbooks repeatedly discuss. In this list, several entities are simply last names, such as *Roosevelt*, which are also ambiguous as to which individual (*Theodore* or *Eleanor*) they refer to. Much of this is due to errors in coreference resolution (see next subsection), such as when the coreference spans across multiple paragraphs of text. To resolve this problem, we pair last names with the most recent full name with that last name that appears beforehand in the text. We also use Wikidata to identify the gender and race of individuals, though White individuals are often missing race/ethnic group labels in this knowledge base, so we manually check these labels as well.

Measuring Space. To accurately measure how often specific people or groups occur in text, we need to also include instances when they are referred to by pronouns like *he* or *she*. To do this we perform coreference resolution, the task of linking textual expressions that refer to the same real-world person. We use the spaCy package and replace pronouns with their full referents. For example, in *Washington's wife, First Lady Martha Washington, attended social events with her husband*, we substitute the pronoun *her* with *First Lady Martha Washington*.

The Clark and Manning (2016) neural coreference model in spaCy was trained on OntoNotes 5.0 (a mix of newswire, broadcasts, and web text) and since textbooks are a different genre, we manually evaluated its performance on a sample of our data (Appendix D). The coreference model achieved a F1 score of .704, with precision = .835 and recall =.618. Our estimated counts of mentions are therefore likely lower than the true number of mentions, but still closer to the true number than if we did not use coreference at all. Another limitation of coreference is that existing models, trained on imbalanced corpora, suffer from gender bias, such as attaching gendered pronouns to nouns referring to stereotypical occupations (Webster et al., 2018). Mitigating these effects is an active area of NLP research.

Research Question 2: How Are Different Groups Described?

After identifying the people discussed in these textbooks, we investigate how they are characterized. Multiple studies using traditional methods have focused on the characterization of women or racial groups in textbooks (Anderson & Metzger, 2011; Blumberg, 2007; Brown & Brown, 2010; Schmidt, 2012). For example, Sarvarzade and Wotipka (2017) looked at the stereotypicality of the women's actions depicted through verbs and visuals in Afghanistan primary school textbooks, and they found that women are often represented as caregivers and mothers.

Here, we demonstrate how relational forms of meaning–associations between words and groups of people–can be identified and extracted using computational methods. The relationships between these words and terms denoting people reveal textbooks' depictions of who people are and what they do.

Identifying Descriptor Words. To extract verbs and adjectives associated with people, we used a part-of-speech tagger and dependency parser, a tool that annotates dependency relations between words (we used a parser by Dozat et al., 2017). This approach is similar to those used by previous work for gathering descriptive attributes of entities in movie plot summaries, books, and news (Bamman et al., 2013; Card et al., 2016; Hoyle et al., 2019). We perform dependency parsing to extract verbs and adjectives associated with people-related terms. Figure 1 illustrates the dependency relations we focus on: adjectival modifier, subject of verbs and object of verbs. In this example, we would extract *individual* and *managed*, since those are two terms associated with *women*.

Comparing Descriptors of Different Groups. To compare the descriptors (adjectives or verbs) of two different groups of people *A* and *B*, we calculate the weighted log-odds-ratio with informative Dirichlet prior of the words associated with them, as described in Section 3.5.1 of Monroe et al. (2008). This method estimates the association of words and groups, building on word frequency counts and an estimate of prior word probability. We chose this method over other, frequentist methods (e.g., difference of proportions, *tf-idf*), because

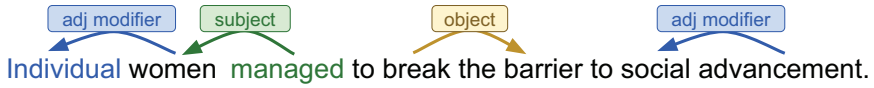

FIGURE 1. *Dependency parsing example.*

it makes use of the prior probability of a word occurring based on counts in a large corpus (in our case, all descriptor words in textbooks), which helps get more accurate signals from words with both very low and very high frequencies. As for the output scores, words with a high positive score are closely associated with Group *A*, while words with a low negative score are associated with Group *B*.

Measuring Connotations of Descriptors. Lexicon-based approaches illuminate the affective and social connotations of words, an area of great importance for the social sciences (Nguyen et al., 2019). This method counts the number of words occurring in a text that are defined in a lexicon as denoting a particular meaning, such as words of positive sentiment. Lexicons have been used since early work in computational content analysis (Stone et al., 1966), and usually have human-generated ratings or labels. Lexicon-based methods are interpretable and computationally inexpensive, but they also have several limitations. They operate under the assumption that the context for which a lexicon is created is similar to the one in which it is applied, which may not hold when a word's meaning varies across contexts (Grimmer & Stewart, 2013). In addition, lexicons contain a fixed number of words and may not always provide good coverage of all relevant words in the corpus (Field et al., 2019).

We apply two families of lexicons: for adjectives, National Research Council (NRC)'s Valence, Arousal, and Dominance (VAD) lexicon (Mohammad, 2018), and for lemmatized verbs (that is, all forms of a verb), the Connotation Frames lexicons of sentiment, power, and agency (Rashkin et al., 2016; Sap et al., 2017). These six metrics we chose to highlight are related to three primary affective dimensions identified in social psychology: power/dominance (strong vs. weak), sentiment/valence (positive vs. negative), and agency/arousal (active vs. passive; Field et al. 2019; Osgood et al., 1957; Russell, 1980). As examples of labeled words in the NRC VAD lexicon, a high valence adjective is *amazing*, a low arousal one is *asleep*, and a dominant one is *competitive*. In the lexicons for connotation frames, X has low agency in the phrase *X obeys*, and for the phrases *X affects Y* and *Y applauds X, X* has power while *Y* does not. In the phrase *X suffered*, the verb *suffered* implies the writer may have positive sentiment toward *X* because it suggests sympathy.

We calculate lexicon scores for social groups following Field et al. (2019), who applied these two lexicon families on online media articles to study portrayals of people in the #MeToo movement (Appendix D). The score for a group of people-related nouns is determined by the average rating of adjectives or verbs describing nouns in that group. We calculate these scores for non-named terms related to different social groups (Appendix Table A1), as well as the top 100 named individuals. We only consider words that have labels in each lexicon, and we use the *z*-score of the calculated values for each lexicon.

Comparing the Association of Words to Different Groups. Similar to previous work looking at gender and ethnic stereotypes (Garg et al., 2018), we also estimate the degree to which certain words are associated with a group by calculating their distance in a latent vector space. We obtain these vector representations, or embeddings, of words and frequent phrases by using a machine learning algorithm that generates them from co-occurrence patterns in corpora. The goal of learning embeddings is to create similar representations for words that occur in similar contexts and different representations for words that occur in different contexts. For example, in history textbooks, the words *women* and *rights* are expected to occur in contexts that are more similar to each other than are the words *women* and *army*, and thus the embedding of *women* should be closer to the embedding of *rights* in the latent space than to the embedding of *army*.

We use the publicly available word2vec skip-gram model (Mikolov et al., 2013) to train our own embedding model on our textbook data. We train our own model instead of using available pre-trained embeddings, in order to capture word co-occurrence patterns present in our textbook data rather than patterns in the dataset (e.g., Wikipedia or the web) used for training the pre-trained model. We use word2vec since it has been shown to be more robust to changes in the data for small datasets (Antoniak & Mimno, 2018) than alternatives (e.g., GloVe; Pennington et al., 2014). We describe the vocabulary and parameter settings in Appendix D.

Semantic similarity between words in the vector space is usually estimated via cosine similarity, which is a measure of how similar the values of a vector are on each dimension. Since word embeddings can be unstable in the case of small corpora, we perform bootstrapping, following Antoniak and Mimno (2018), to ensure that we have robust estimations of word similarity (Appendix D).

To identify words for our analyses, we select themes that are relevant to previous studies on the representation of gender in history textbooks, such as the home, the workplace and politics (Sarvarzade & Wotipka, 2017; Schmidt, 2012). We match these three themes with the *home, work* and

achievement word categories in the Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker et al., 2015), respectively. For each category, we select the words in LIWC that are the most frequent in our data and filter out words that, in our domain, are unlikely to be used in a sense that fits their LIWC category (e.g., *house* in the context of history books is usually used to refer to a political institution rather than a synonym of *home*). Then, for each term (e.g., *household*) we calculate the mean cosine similarity between that term and all the terms referring to a particular gender (e.g., *she, her, woman*, etc.).

Research Question 3: What Are Prominent Topics and How Are They Related to Groups of People?

Next, we move from methods that draw out word-to-word relationships in text to methods that enable the study of word-to-topic and topic-to-topic ones. Textbook researchers often study the prevalence of topics and the relationship among them, as they can shed light on the perspective and framing that a particular textbook promotes. The words associated with each topic and the way in which the topics are related, however, is usually left to the coder to define (see Lachmann & Mitchell, 2014, for an example where the authors use hand-curated word categories). This may lead to low interrater reliability due to annotation bias or coding error from overlooking relevant items. Computational methods built on word co-occurrence patterns allow for the automatic grouping of words into topics, thereby potentially uncovering relational meanings with greater efficiency than manual coding when attempting to analyze large amounts of text. Subject area expertise remains of central importance, as researchers must thoughtfully attribute meaning to the automatic groupings.

Identifying Topics. Topic modeling is a central approach for automatically discovering topics in a collection of documents. There are several different kinds of topic models, the most commonly used being Latent Dirichlet Allocation (LDA; Blei et al., 2003). LDA represents the distribution of topics within documents and the distribution of words within each topic. Such LDA models have been previously applied to an enormous variety of texts and genres (Boyd-Graber et al., 2017). In educational contexts, LDA models have been used to analyze student writing (Chen et al., 2016) and MOOC (Massive Open Online Courses) discussion forums (Ramesh et al., 2014; Reich et al., 2015; Vytasek et al., 2017). We employ LDA to study the prominence of different topics within and across textbooks and the prominence of words related to different groups of people within and across topics.

Topic models require a collection of documents as input. We perform topic modeling at the sentence level, which provides us with a large number of similarly sized documents (17 tokens on average) that are suitable for inducing stable estimates of a wide range of topics. To build our vocabulary for the model, we first remove function words (e.g., *the, it, have*) based on a list of stopwords included in MALLET, an off-the-shelf tool for topic modeling (McCallum, 2002). We also perform stemming[5] via the SnowballStemmer (Porter, 2001). We use our resulting set of tokens (unigrams and bigrams) to compile document-to-token counts, which serve as an input to MALLET. We build a topic model with $k = 50$ topics. We expect there to be a large number of different topics in the textbooks, but we limit the number of topics to 50 because in experiments with more topics ($k = 75$, $k = 100$, $k = 300$), we found that the topics were too fine-grained for our analyses (e.g., multiple topics representing multiple wars). Depending on the research focus, a lower or higher number of topics may prove necessary. We explain other parameter settings and decisions in Appendix D.

To understand which topics relate to which social groups, we can look at the topics in which non-named people-related terms have a high probability. We consider a topic to be associated with a term if the term is among the top ten highest probability terms for that topic. Since we remove function words, we expect high probability words in each topic to represent a collection of semantically related words, also known as a semantic field. Thus, the more topics a term is associated with, the more semantically diverse we expect the discussion around that term to be–henceforth, we refer to this phenomenon as topical diversity.

Comparing the Prominence of Topics Across Books. We estimate the prominence of a given topic within a textbook by taking the mean probability of the topic across sentences in that book. We measure the prominence of multiple topics (henceforth, a topic group) associated with a term by summing their average probabilities. We calculate relative topic prominence of a topic group pair by calculating the ratio of their prominence within a book. We compare the relative prominence of topic group pairs across books instead of the prominence of a single topic group because the former method is more robust to noise arising from different textbooks having different lexical distributions and hence, topic probabilities. We remove three books that only cover half of U.S. history from topic-related analyses.

Results and Discussion

Research Question 1: How Much Space Is Allocated to Different Groups?

Identifying People-Related Terms. The three most common nonnamed people terms overall are *people, women*, and *his*. The high frequency of *his* suggests that some pronouns were not resolved with the noun they refer to during coreference resolution. Most terms are unmarked by gender or race/ethnicity, though from the percentage of those that are, men are mentioned more often than women (Figure 2). Black people

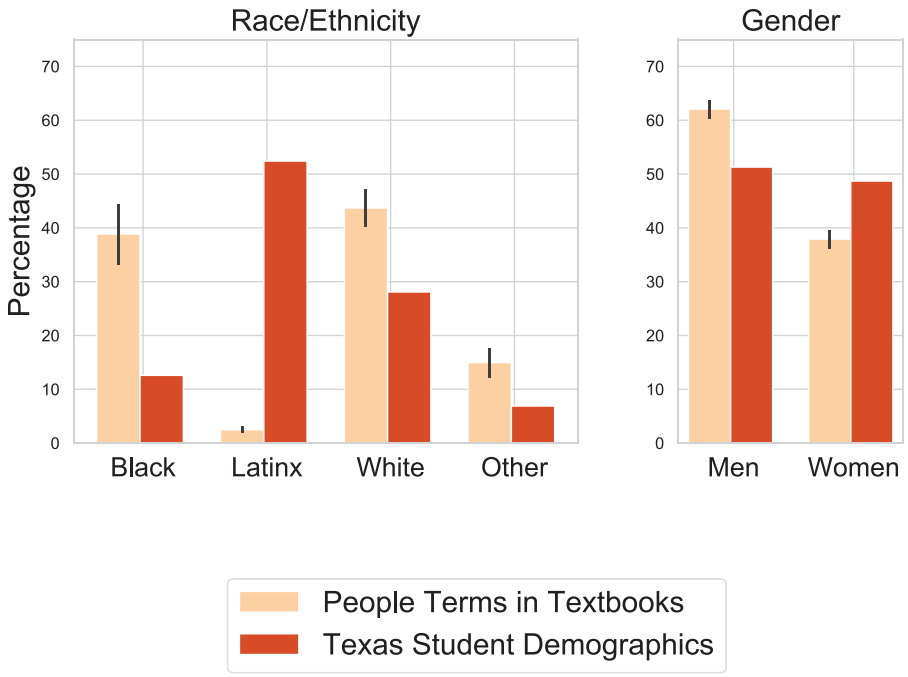

FIGURE 2. *Percentage of people of different demographics in Texas textbooks and schools.*
Note. Error lines for nonnamed people terms show 95% confidence intervals across textbooks. There are no error lines on the bars for Texas students, which is an aggregate of all schools' student demographics drawn from National Center for Education Statistics.

are the most common nonwhite racial/ethnic group discussed (Figure 2). Though a slight majority of people marked by race/ethnicity are nonwhite, it does not mean these textbooks do not focus on the history of White Americans; rather, many terms (e.g., *pioneer, farmer, priest*) seem to implicitly convey or assume Whiteness. This is a common kind of "reporting bias," in which people are less likely to state the most common properties of an entity, since they believe the audience will assume the majority demographic as the default (Gordon & Van Durme, 2013).

Perhaps the most striking finding with regards to the ethnicities of people mentioned in textbooks is the scarcity of Latinx groups (Figure 2). Previous work has shown the importance of culturally relevant education, such as students seeing their personal identities represented in school curriculum, in improving students' learning outcomes (Aronson & Laughter, 2016; Dee & Penner, 2017). However, despite the fact that demographic data shows 52.42% of students in Texas are Latinx, they are only mentioned 961 times across all textbooks, which accounts for only 0.248% of people terms and 2.23% of people terms marked by ethnicity/race. Latinx groups tend to be discussed in coverage of the Mexican-American War, as well as in contrast to incoming White settlers: *[Early pioneers] left the Oregon Trail . . . and mostly settled in the interior along the Sacramento River, where there were few Mexicans* (Bedford *America's History*, Henretta et al., 2014, p. 413). Indigenous peoples and Asian Americans are also scarce in the texts. We do not expect to see representation that is directly proportional to population demographics, but the distinctions we find provide empirical information for future research on how and why curricula shape students. Given most research in this field relies on hiring human coders for the task of identifying social groups in text (e.g., Bromley et al., 2011), we examined how our NLP results compare with the traditional approach (Appendix E).

Identifying Famous People. The most frequently mentioned named people across textbooks are almost entirely White men in politics (Table 2). Our books place a significant amount of focus on these named individuals; one fifth of sentences mention at least one of the top 50 named people. The only woman in the top 50 is First Lady Eleanor Roosevelt, who is the 28th most common person discussed. This finding agrees with prior work, which also found that Eleanor Roosevelt is the most mentioned woman in U.S. history textbooks (Tetreault, 1986). In our textbooks, the next most common woman is American activist Jane Addams, ranked 54th. The limited number of people of color within the top 50 include President Barack Obama (29th), activist Martin Luther King, Jr. (30th), slave Dred Scott (42th), and abolitionist Frederick Douglass (44th), who are all Black. Thus, the amount of space allotted for famous people featured in history textbooks is dominated by a single demographic, with a few exceptions. This result may be a consequence of the textbooks' focus on politics rather than everyday

TABLE 2
The Top 30 Most Common Named People Across All Textbooks

Name	No. of appearances	Wikidata gender
Andrew Jackson	3,347	Male
Thomas Jefferson	3,033	Male
Franklin Delano Roosevelt	2,672	Male
Richard Nixon	2,659	Male
Theodore Roosevelt	2,627	Male
Ronald Reagan	2,294	Male
John F. Kennedy	2,176	Male
Lyndon Johnson	1,546	Male
George W. Bush	1,291	Male
Woodrow Wilson	1,269	Male
Alexander Hamilton	1,234	Male
Harry S. Truman	1,227	Male
Bill Clinton	1,211	Male
James Madison	1,173	Male
John Adams	1,156	Male
Andrew Johnson	1,125	Male
Robert E. Lee	1,053	Male
Abraham Lincoln	968	Male
Adolf Hitler	961	Male
George Washington	875	Male
Eisenhower[a]	856	(None)
Ulysses S. Grant	803	Male
John Quincy Adams	789	Male
Jimmy Carter	785	Male
John Brown	694	Male
Herbert Hoover	660	Male
George H. W. Bush	658	Male
Eleanor Roosevelt	573	Female
Barack Obama	566	Male
Martin Luther King Jr.	563	Male

Note. The names are obtained after Wikidata name standardization, frequency filtering, and last name disambiguation.
[a]Because *Eisenhower* did not manage to be automatically disambiguated and is not a full name, Wikidata does not have a gender label for it, but this name most likely refers to the White president Dwight D. Eisenhower, who is a man.

life and sociocultural movements–a phenomenon that we return to in our later results.

Link to External Factors. It is well established that the creation of textbooks is deeply political (Apple, 1992; Apple & Christian-Smith, 2017; Foster, 1999). We found that percentages of nonnamed women and Black people in textbooks are positively correlated with the median percentage of Democratic votes in counties that purchased each textbook (Figure 3). The Pearson correlation r between the percentage of mentions of Black people and the percentage of democratic votes during the 2016 presidential election is .519 ($p < .05$) and between the percentage of women mentions and Democratic votes is .583 ($p < .03$). In Pearson's *U.S. History*, which was purchased in the most Republican counties, 1.82% of nonnamed people mentions are Black and 4.87% are women, while in *Give Me Liberty*, which was purchased in the most Democratic counties, these values are 8.58% and 6.82%, respectively. State-adopted textbooks or textbooks such as Jarrett's *Mastering the TEKS* that adhere to Texas-specific standards in particular have less representation of Blacks and women and are used in more conservative counties (Figure 3). The percentage of Latinx mentions did not show any significant trend (Pearson $r = -.107, p = .703$), likely due to the low prevalence across all textbooks (variance $\sigma^2 = .01$). While emphases on diversity are quite low across all districts, there are significant differences in district purchasing, with districts in more conservative counties using less diverse books.

Research Question 2: How Are Different Groups Described?

Comparing Descriptors of Different Groups. A log odds comparison of the words associated with Black people with those associated with Whites and people terms unmarked for ethnicity reveals that Black people tend to be described with words related to slavery, such as *free* and *runaway*, and not words related to politics such as *political* and *federal* (Figure 4). We also compared the words associated with women with those associated with men and terms unmarked for gender. Women tend to be described with words related to their marital status, and not with words related to the military or government, which is consistent with other stereotypical portrayals of women in media (Collins, 2011). These results are also consistent with the historical exclusion of nonmen and non-White people from politics, and further illustrate the kinds of contexts in which these social groups are portrayed.

Measuring Connotations of Descriptors. We used lexicons to categorize descriptors associated with different groups of people and famous individuals. Though the concepts labeled in these lexicons, such as dominance in NRC VAD, could be interpreted as positive attributes for people to have, these labels do not advocate for how people should be described in textbooks. In the first lexicon for connotation frames, 85.0% of a total of 165,386 non-unique verbs (3,983 are unique) attached to people in these textbooks had sentiment labels, 92.4% had power labels, and 61.1% had agency labels. Our second lexicon, NRC VAD, contains 68.8% of the 108,033 nonunique adjectives (7,563 are unique) describing people. A line of future work would be to induce scores for words not labeled in these lexicons and customize their scores for history textbooks.

Our analysis of verbs using the lexicon for connotation frames showed that Black people are depicted with less power and agency than other social groups (Figure 5).

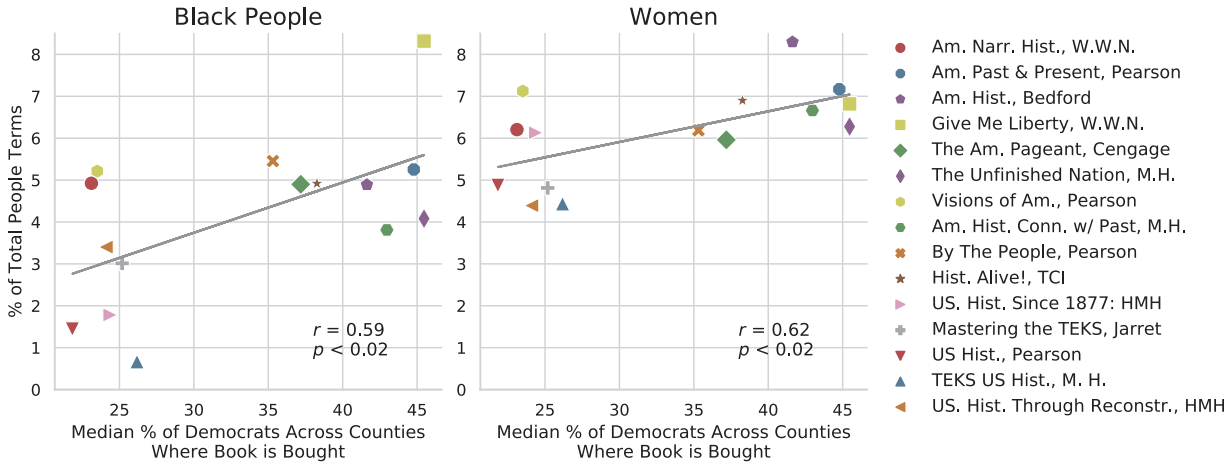

FIGURE 3. *Political factors and textbook distribution versus percentage of mentions.*
Note. State-adopted textbooks are represented by triangles.

These differences are largely due to their appearance in the context of slavery and racial oppression, and they are the object of high-power verbs such as *owned* and *barred*. This finding contrasts with new historical research that emphasizes the power and agency of Black people in freeing themselves from enslavement and oppression (Devlin, 2018; Hines, 2016). Named individuals have the highest agency and power, performing political actions such as *veto* and *initiate* (Figure 5). Additionally, the sentiment of a writer toward a subject or object is most positive for women (Figure 5). Examples of common verbs associated with women that have high positive sentiment scores include *marry* and *help*. This result illustrates the importance of examining the actual words involved in the calculation of a lexicon-based score, as lexicons may have labels that gloss over words' complex and context-dependent connotations. Due to the small amount of Latinx representation in textbooks, the confidence intervals for their words' lexicon scores are large, and no clear conclusions can be drawn about the words associated with them.

Our analysis of adjectives using the second lexicon, NRC VAD, reveals a few trends that complement our verb-based findings. For example, the adjectives describing Black people, such as *slave* and *inferior*, have lower dominance ratings than those describing other groups. Additionally, named entities tend to be described with high arousal adjectives, such as *worried, victorious*, and *furious*.

Comparing the Association of Words With Different Groups. Another indicator that women are associated with domestic activities in textbooks can be seen in our word embedding results. First, the most similar tokens in the textbook embedding space to words denoting women (*woman, women, female, she, her, hers*), as measured by cosine similarity, are words and phrases related to the domestic sphere. These tokens are *woman's husband* (.58), *wife and mother* (.57), *housewife* (.56), *breadwinner* (.56), *husband* (.54), where parenthetical values indicate cosine similarity estimated via bootstrapping. Second, by using terms within LIWC categories, we find that men (*man, men, male, he, him, his*) are less closely associated with the home and more closely associated with achievement than women (Figure 6).

Women are also more closely associated with work-related terms than men. This result is consistent with that of Schmidt (2012) who found that the greatest number of references to women occur in the context of the workplace in recent U.S. history curricula. However, again following Schmidt (2012), the strong association between women and the workplace does not imply the textbooks take a feminist view. The degree to which women's agency and their *choice* (rather than *need*) to work is emphasized, and the type of jobs that they are associated with is also an important part of this framing. Exploring these aspects could be a useful contribution of future work.

Research Question 3: What Are Prominent Topics and How Are They Related?

Identifying Topics. Table 3 shows the highest probability terms for the 10 most prominent topics that emerge from the texts. Topics are ordered by their average probability across all books (see all 50 topics in Appendix Table B1). Note that the topic probabilities are similar, which is expected given that we keep the prior probabilities of the topics and the words fixed (Appendix D).

In topic modeling, careful interpretive sense-making is required as the researcher determines the label or meaning of word groupings. Examining the 10 highest probability words for all 50 topics manually, we find that seventeen topics are associated with formal politics (including stems such as *govern, presid, polit, federal*), two with social movements

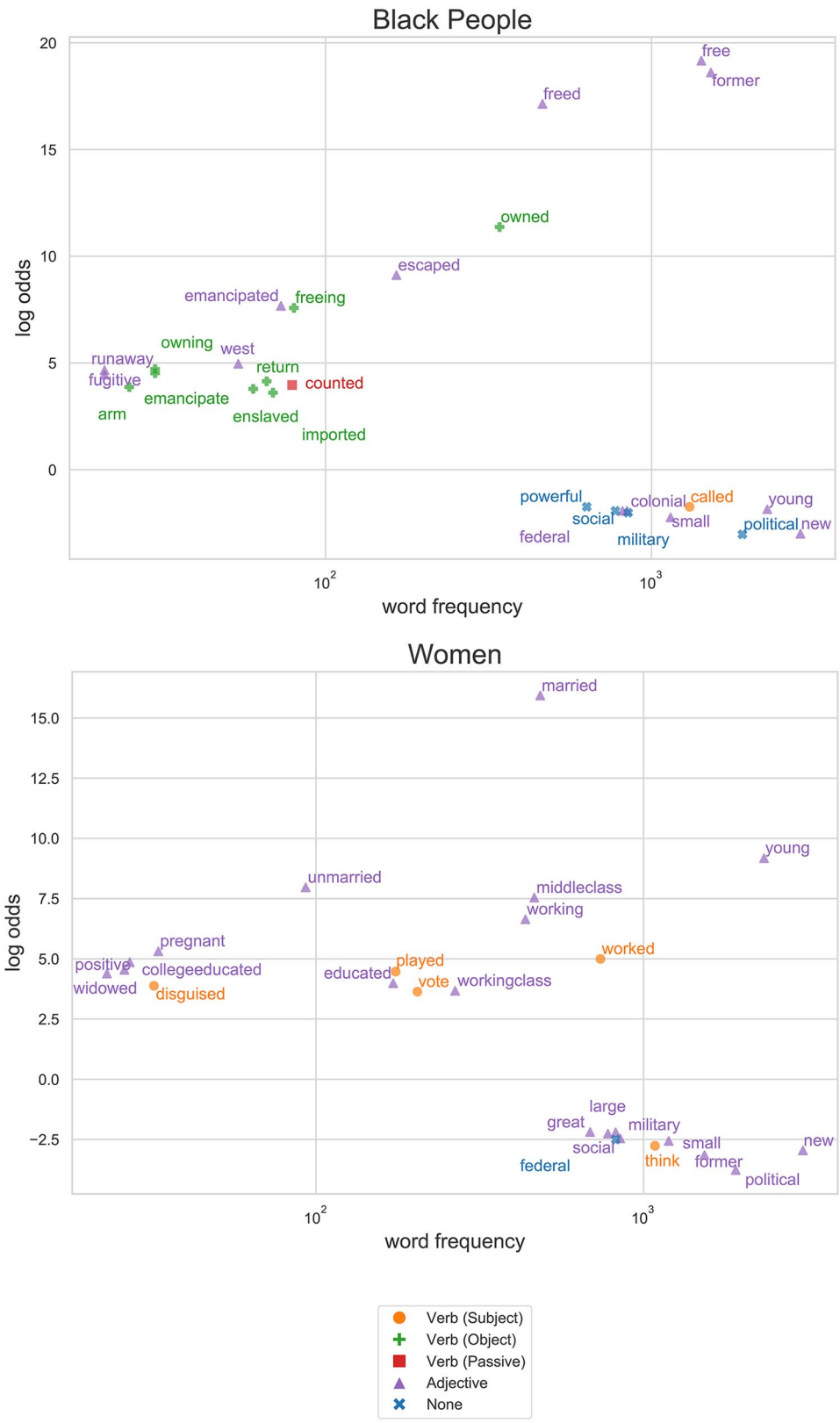

FIGURE 4. *Log-odds-ratio of words associated with Black people and women.*
Note. These plots only show descriptors that occur at least 20 times. A point's color corresponds to the most common way the word relates to the people or person being described, and "None" means that word does 'not ever co-occur with that social group. Words above the 0 line are discussed more often in reference to Black people/women while words below the line are discussed more often in reference to White people and people terms unmarked for ethnicity.

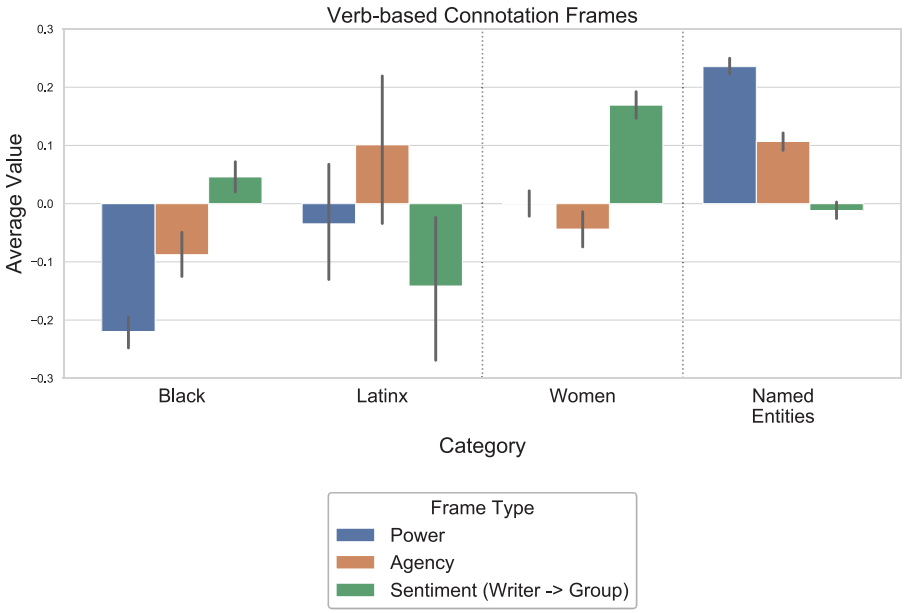

FIGURE 5. *Verb-based connotation frames of power, agency, and sentiment for social groups.*
Note. Error bars show 95% confidence intervals. "Named Entities" includes the top 100 named entities after Wikidata name standardization. The Power bar for Women looks empty because its average value is .01, or close to 0.

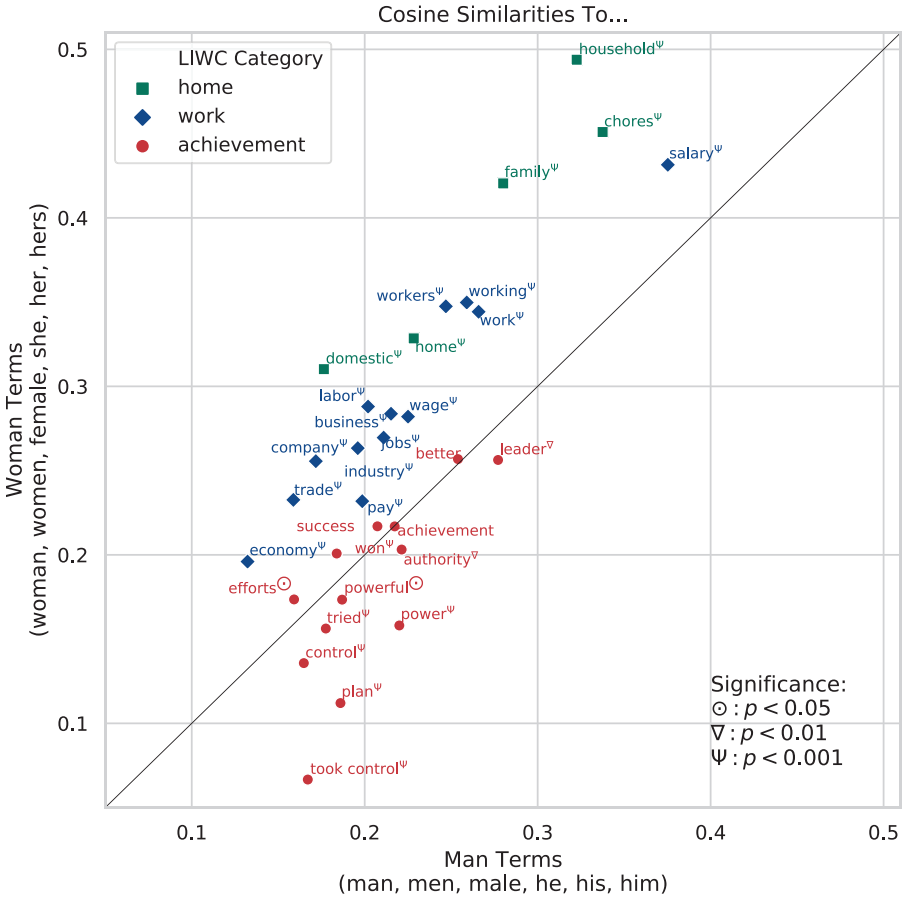

FIGURE 6. *Cosine similarity of gendered terms and words related to home, work, and achievement.*
Note. Significance for each term is calculated via a two-tailed *t* test. Words above the 45° line are discussed more often in reference to women and words below the line are more often linked to men.

TABLE 3
The 10 Most Prominent Topics in Our Data

No.	Topic Probability	Top 10 topic terms
1	.022	armi, general, confeder, troop, union, forc, command, battl, british, victori
2	.0218	democrat, parti, republican, elect, vote, candid, won, voter, major, popular
3	.0213	read, inform, sourc, newspap, write, book, chapter, map, publish, learn
4	.0213	man, hand, boy, thing, back, day, eye, told, cloth, dress
5	.0211	centuri, industri, chang, growth, develop, economi, econom, revolut, region, increas
6	.021	european, north, america, spanish, explor, empir, europ, trade, spain, africa
7	.021	water, river, cattl, miner, mountain, gold, mine, food, west, forest
8	.0209	unit, war, world, state, nation, civil, end, power, america, year
9	.0208	explain, identifi, role, describ, effect, event, analyz, play, import, impact
10	.0206	german, germani, soviet, alli, franc, soviet union, europ, hitler, russia, unit

Note. Topics are ordered by their average probability across textbooks.

TABLE 4
Topics Associated With Different Groups of People

Terms referring to groups	Topics
women, woman	• movement, women, organ, group, civil right, right, leader, african, polit, equal • men, women, famili, children, young, work, woman, home, mother, husband
man, men	• soldier, thousand, die, kill, hundr, death, year, day, men, fight • human, natur, man, person, thing, moral, reason, believ, good, individu • man, hand, boy, thing, back, day, eye, told, cloth, dress • men, women, famili, children, young, work, woman, home, mother, husband
white	• indian, nativ, land, tribe, west, settler, american, white, western, frontier • african, black, slave, white, southern, free, south, american, slaveri, northern
black, african american	• african, black, slave, white, southern, free, south, american • king, march, day, protest, washington, demonstr, polic, martin luther, mob, black
native american	• indian, nativ, land, tribe, west, settler, american, white, western, frontier
hispanic, latinx, mexican	• mexican, mexico, unit, texa, california, territori, spanish, florida, spain, claim

Note. We define association between a term and a topic as the term occurring in the 10 highest probability words for the topic. Note that the same topic can represent multiple groups.

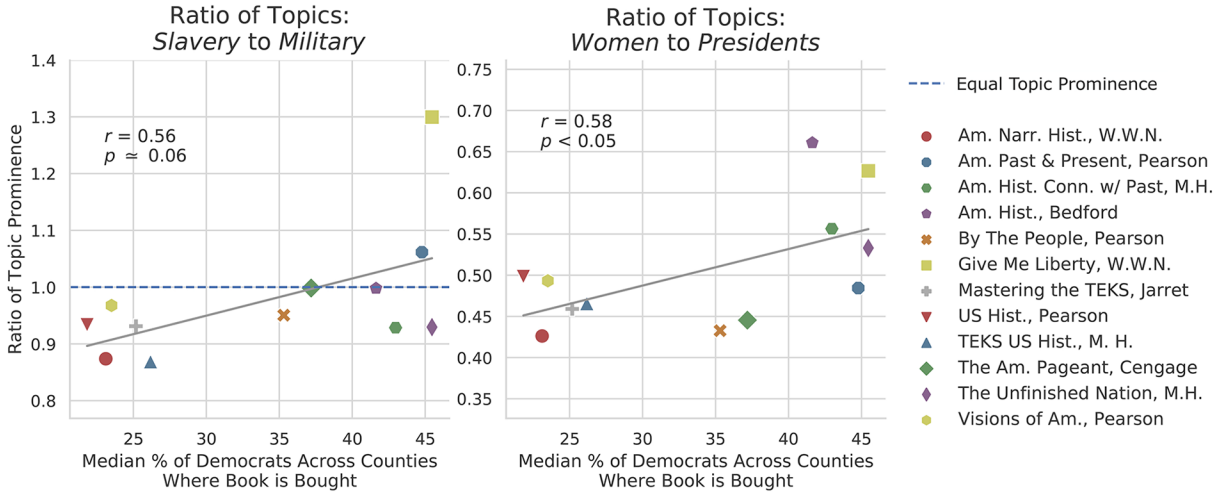

FIGURE 7. *Correlation of topic ratios with political orientation of counties where books are purchased.*
Note. Books represented by triangles are state adopted. A score of 1 means that topics are equally discussed, a score above 1 means that slavery is discussed more than the military, or that women are discussed more than presidents.

(*movement, protest, civil*), and three with everyday workers (*farmer, worker, soldier*). Since each topic occurs with a relatively similar probability (1.8%–2.2%), the greater number of topics provides evidence that formal politics are emphasized more than topics that focus on the voices of the citizens, in contrast to recommendations by some scholars (e.g., Loewen, 2008; Zinn, 1984).

To dig deeper into our topics, we consider how they are distributed across the different non-named people groups we identified earlier. We study the topical diversity of groups by looking at the number of topics they are associated with (Table 4). As for gender, we find that *woman/women* are associated with two topics–one related to social movements (women's rights) and the other to family. *Man/men* are also associated with the family topic, as well as with three others: one related to the military, another to decision making/morality, and another to mentions in quotes (Appendix F). As for ethnic groups, we find that white only occurs in the context of other ethnicities, suggesting that Whiteness is unmarked unless it is contrasted with minority ethnicities. Black people are associated with two topics, one related to slavery and the other to civil rights, Native Americans are discussed in the context of settlers, and Latinx people in the context of territorial claims. These topics suggest that the discussion of minority ethnicities is dominated by topics where the relationship of the minorities to the majority group is highlighted in some way. These topics, combined with further qualitative analyses, could allow for a better understanding of textbooks' degree of multifocality (Gordy & Pritchard, 1995).

Comparing the Prominence of Topics Across Books. We also take a closer look at the relative prominence of topics that have been studied in previous research on U.S. history textbooks due to their relevance to the representation of people and events (Anderson & Metzger, 2011; Schmidt, 2012). For example, we compare the prevalence of discussions of slavery with the military in textbooks, and the prevalence of discussions of women relative to discussions of presidents. As Figure 7 illustrates, we find that books that are purchased in more Republican counties tend to talk more about the military (topics associated with *armi, militari*) than about slavery (*slave, slaveri;* $r = .56$, $p \approx .06$). We also find a positive correlation between the median percentage of Democrats where books are purchased and the relative prominence of topics associated with women (*women, woman*) versus ones related to *presidents* ($r = .58$, $p < .05$). Nonetheless, despite differences across books, all books talk more about presidents than women. In fact, both of our results on between-book differences in their representation of people (Figures 3 and 7) suggest that the between-book variation is small compared with pervasive similarities reflecting a deeper, shared historical narrative that is conveyed in these books.

Conclusion

Textbook research, and other fields where traditional content analysis methods have dominated, are particularly fruitful settings for the application of NLP methods. Computational tools not only allow for faster, more comprehensive, and bigger studies than prior research but can also enable different insights. They can illuminate the scope and scale of discursive trends in new ways and deepen understandings about the meaning of concepts through the co-occurrence of words, people, and topics. Furthermore, these quantitative measures combined with larger sample sizes facilitates analyzing links between the text and external influences.

In our work on U.S. history textbooks used in Texas, NLP methods for identifying people reveal that Latinx people are virtually absent from textbooks and named individuals are mostly white men. Measured associations between words show that women are mentioned in the contexts of marriage, home, and work, and Black people are involved in actions with low agency and power. Topic modeling demonstrates that books focus more on political history than social history, and discussions of minority ethnicities center on their relationships with White people. We also find that more conservative counties tend to purchase textbooks with less representation of marginalized groups, but that the systematic variation across textbooks is small relative to their pervasive similarities.

Future methodological work would be to develop novel algorithms, models, and lexicons specifically for the domain of social science textbooks, as many of the methods we demonstrate were previously applied to other domains such as news, social media, or fiction. In addition, echoing our introduction, this type of work is inherently interdisciplinary, which means computational approaches cannot operate alone. In-depth qualitative analyses based on the expertise of education researchers and other social scientists still remain crucial not only for a thorough understanding of textbook content but also for interpreting and contextualizing the computational results themselves. The methods are also still under constant development, including efforts to improve fairness (Hovy & Spruit, 2016).

Our contribution centers on methods for describing what textbooks contain. We hope that our approaches can support further research, such as that of Apple and Christian-Smith (2017), on explaining the mechanisms that lead books to contain certain content. For example, one could delve deeper into our preliminary association between political environment and content, by looking at the composition of a state's board of education or other factors. We hope more nuanced measures will inform discussions about how textbooks can be improved, understanding that norms are constantly evolving and there may not be a single correct answer to what should be in a book. Used thoughtfully, NLP and other methods linked to the rise of artificial intelligence and data science have the potential to generate novel conceptual insights for education research, policy, and practice.

Appendix A

TABLE A1
List of Categorized People-Related Terms in Their Singular Noun Form

Black	Latinx	Other minority	White	Women	Men
slave (9,339)	mexican (523)	immigrant (3,993)	white (6,350)	woman (14,718)	man (11,558)
black (5,673)	latino (136)	tribe (1,657)	colonist (3,172)	her (2,181)	his (9,902)
african (476)	hispanic (98)	indian (1,394)	british (2,224)	wife (1,502)	he (6,536)
enslaved (437)	mexican-american (46)	minority (897)	english (1,350)	mother (1,132)	king (1,181)
freedman (336)	bracero (44)	native (871)	european (1,038)	she (954)	husband (1,044)
negro (330)	[puerto] rican (33)	japanese (372)	spanish (856)	girl (746)	him (980)
fugitive (311)	chicano (30)	refugee (289)	french (818)	female (547)	son (964)
african-american (278)	mexicano (25)	jewish (246)	german (652)	daughter (436)	father (881)
runaway (275)	tejano (18)	vietnamese (209)	slaveholder (407)	feminist (308)	himself (847)
ex-slave (101)	panamanian (6)	foreigner (158)	puritan (206)	sister (298)	male (774)
freeman (71)	latina (2)	savage (156)	italian (196)	lady (263)	boy (680)
n*gger (55)		asian (105)	portuguese (104)	widow (171)	brother (585)
freedperson (47)		filipino (102)	slaveowner (90)	queen (159)	congressman (400)
mulatto (46)		iraqi (87)	minstrel (89)	witch (154)	uncle (387)
creole (30)		korean (86)	secessionist (64)	lesbian (134)	businessman (332)
freedwoman (10)		cherokee (85)	nazi (57)	mistress (130)	cowboy (275)
africanamerican (2)		iroquois (82)	yankee (50)	flapper (91)	gentleman (206)
		tribal (78)	conqueror (46)	herself (78)	militiaman (168)
		outsider (76)	ex-confederate (43)	housewife (65)	lord (146)
		ethnicity (71)	anglo (38)	grandmother (61)	spokesman (134)
		muslim (65)	vice-president (36)	bride (57)	emperor (130)
		Taztec (60)	anglo-american (30)	midwife (55)	policeman (112)
		pueblo (60)	greek (27)	seamstress (52)	statesman (107)
		sioux (59)	jesuit (23)	heroine (44)	clergyman (106)
		non-white (47)	roman (23)	aunt (32)	grandfather (85)
		nonwhite (45)	hungarian (17)	princess (30)	chairman (77)
		iranian (45)	austrian (15)	waitress (28)	serviceman (73)
		caribbean (40)	anglo-saxon (13)	laundress (27)	frontiersman (71)
		shawnee (34)	czech (13)	goddess (23)	brethren (67)
		cheyenne (29)	tory (12)	niece (19)	seaman (62)
		inca (29)	klansman (10)	nun (18)	countryman (62)
		jew (25)	irishman (8)	hers (13)	minuteman (50)
		subculture (24)	englishman (8)	cowgirl (0)	grandson (46)
		apache (24)	australian (7)	bridesmaid (0)	guy (45)
		navajo (22)	amish (6)		workingman (45)
		anasazi (22)	scandinavian (5)		foreman (45)
		israeli (21)	moravian (4)		workman (39)
		japanese-american (21)	latin (3)		salesman (38)
		mohawk (19)	victorian (3)		horseman (35)
		algonquian (19)	georgian (3)		nobleman (32)
		seminole (19)	bolshevik (2)		cattleman (32)
		palestinian (18)	slav (2)		friar (32)
		gypsy (18)	spaniard (2)		fisherman (27)
		jamaican (15)	anglo-texan (2)		widower (26)
		huron (13)	briton (2)		forefather (25)
		powhatan (13)	englander (1)		journeyman (25)
		perce (12)			patriarch (23)

(continued)

TABLE A1 (CONTINUED)

Black	Latinx	Other minority	White	Women	Men
		comanche (12)			tradesman (21)
		dakota (12)			fireman (21)
		marginalized (10)			gunman (20)
		arapaho (10)			prince (20)
		choctaw (8)			nephew (19)
		haitian (7)			rifleman (14)
		chickasaw (7)			guardsman (13)
		barbadian (6)			pope (9)
		lakota (6)			duke (5)
		sauk (6)			groom (4)
		hebrew (5)			dairyman (1)
		mandan (4)			
		mexica (4)			
		asian-american (2)			
		hopi (2)			
		puebloan (1)			

Note. These nouns were manually filtered from all heads of noun phrases across textbooks, and the frequency in brackets also includes occurrences where they are used as adjectives to mark other people-related nouns, e.g., *Black man*. Our analyses included an additional 1,665 terms, such as *worker* and *village*, that were not categorized into a demographic group.

Appendix B

TABLE B1
All 50 Topics Used in Our Topic Modeling Analysis

Topic probability	Top 10 topic terms
.022	armi, general, confeder, troop, union, forc, command, battl, british, victori
.0218	democrat, parti, republican, elect, vote, candid, won, voter, major, popular
.0213	read, inform, sourc, newspap, write, book, chapter, map, publish, learn
.0213	man, hand, boy, thing, back, day, eye, told, cloth, dress
.0211	centuri, industri, chang, growth, develop, economi, econom, revolut, region, increas
.021	european, north, america, spanish, explor, empir, europ, trade, spain, africa
.021	water, river, cattl, miner, mountain, gold, mine, food, west, forest
.0209	unit, war, world, state, nation, civil, end, power, america, year
.0208	explain, identifi, role, describ, effect, event, analyz, play, import, impact
.0206	german, germani, soviet, alli, franc, soviet union, europ, hitler, russia, unit
.0204	cultur, societi, american, tradit, life, group, immigr, distinct, valu, reflect
.0204	presid, kennedi, johnson, reagan, nixon, administr, truman, polici, eisenhow, bush
.0203	popular, show, imag, paint, artist, photograph, depict, music, televis, audienc
.0202	human, natur, man, person, thing, moral, reason, believ, good, individu
.0202	african, black, slave, white, southern, free, south, american, slaveri, northern
.0202	indian, nativ, land, tribe, west, settler, american, white, western, frontier
.0201	social, reform, polit, progress, econom, societi, interest, class, effort, system
.0201	peopl, freedom, liberti, equal, american, countri, free, democraci, idea, great
.0201	slaveri, southern, lincoln, northern, union, south, territori, free, north, compromis
.0201	debat, conflict, tension, polit, issu, continu, divis, cold war, era, controversi
.0201	movement, women, organ, group, civil right, right, leader, african, polit, equal
.0201	percent, million, number, popul, year, immigr, increas, half, larg, total
.02	constitut, right, amend, vote, citizen, convent, deleg, state, bill, congress

(continued)

TABLE B1 (CONTINUED)

Topic probability	Top 10 topic terms
.02	face, problem, depress, econom, suffer, economi, caus, great depress, crisi, prosper
.0198	bank, money, tax, pay, debt, loan, rais, fund, paid, govern
.0198	worker, labor, work, union, job, employ, strike, factori, industri, wage
.0198	govern, power, feder, nation, peopl, author, constitut, state, system, unit
.0198	roosevelt, wilson, peac, presid, treati, negoti, theodor roosevelt, taft, leagu, agreement
.0197	men, women, famili, children, young, work, woman, home, mother, husband
.0197	citi, york, urban, hous, live, town, center, communiti, move, chicago
.0197	railroad, build, line, technolog, transport, road, develop, travel, invent, canal
.0197	good, trade, product, manufactur, market, import, produc, economi, consum, tariff
.0196	farmer, farm, planter, small, land, cotton, plantat, crop, famili, larg
.0196	jackson, jefferson, adam, federalist, support, presid, hamilton, andrew jackson, whig, republican
.0196	soldier, thousand, die, kill, hundr, death, year, day, men, fight
.0196	king, march, day, protest, washington, demonstr, polic, martin luther, mob, black
.0196	vietnam, forc, militari, troop, south, unit, war, attack, iraq, communist
.0195	ship, japanes, japan, island, china, navi, british, sea, chines, attack
.0195	mexican, mexico, unit, texa, california, territori, spanish, florida, spain, claim
.0194	crime, charg, prison, accus, trial, critic, convict, public, communist, murder
.0194	religi, church, christian, protest, god, minist, cathol, religion, puritan, communiti
.0193	program, deal, feder, provid, public, creat, administr, aid, roosevelt, work
.0193	english, england, virginia, coloni, pennsylvania, settler, governor, establish, york, dutch
.0193	british, colonist, coloni, french, britain, independ, king, revolut, parliament, patriot
.0192	suprem court, court, decis, law, rule, case, justic, legal, constitut, separ
.0192	compani, busi, railroad, corpor, industri, steel, product, manag, oil, trust
.0191	offic, presid, hous, senat, elect, member, repres, appoint, congress, serv
.0191	act, pass, law, congress, legisl, bill, immigr, feder, reconstruct, prohibit
.0191	john, william, wrote, jame, son, henri, georg, name, smith, brown
.0186	school, educ, public, student, univers, colleg, servic, high, children, train

Note. Topics are sorted by their average probability across books.

Appendix C

Textbook Selection

We select textbooks based on district-level purchase data documented by the Texas Education Agency. The data include disbursements that occurred between 2015 and 2016 and requisitions that occurred between 2015 and 2017. Requisitions are entirely state-adopted textbooks, while disbursements can be both state-adopted and nonstate-adopted. Requisition data are already organized by standardized titles, authors, and publisher names. We manually disambiguate this information for disbursements, ignoring cases where the listed title is too generic, such as "History". We filter the books to include only titles that appear in more than 10 district-level transactions, where each transaction represents an entry in the original disbursement and requisition data. Texas's nonelementary U.S. history curriculum is segmented so that the first half of material is usually taught in 8th grade and the second half is taught in high school. Thus, when books exist as two separate volumes, we combine them into one. As not all transactions listed the specific edition or publication date of a textbook, we obtained the most recent editions published before 2017 that were available for purchase online. Table C1 lists all textbooks we use, with additional information on the books.

Textbook Content

Each textbook is chronologically ordered and generally covers United States history starting with European exploration of the Americas and colonization, and ends with the present day. If a textbook consists of two volumes, they are split by the Civil War and Reconstruction period. In many textbooks, chapters that tackle individual themes are grouped into larger units that correspond to time periods. For example, in *Give Me Liberty: An American History*, the unit covering the Gilded Age (1870–1890) contains separate chapters for politics and life in the American West. Since Texas history is taught in a different grade-level separately from U.S. history, the Texas editions of U.S. history textbooks do not focus on Texas history but adhere to Texas state standards for curriculum.

TABLE C1
The U.S. History Textbooks Used in Our Study, With Publication Information

Title	Author/s	Publisher	Publication date	State adopted?	Number of purchases
United States History: Early Colonial Period Through Reconstruction, Texas Edition [ISBN: 978-0-544-32028-4]	Deborah G. White, William Deverell	Houghton Mifflin Harcourt	2016	Yes	501
The Americans: United States History Since 1877, Texas Edition [ISBN: 978-0-544-32140-3]		Houghton Mifflin Harcourt	2016	Yes	426
TEKS United States History to 1877 (I) & *Since 1877* (II) [ISBNs: 978-0-076-59810-6, 978-0-076-60854-6]	Alan Brinkley	McGraw-Hill	2016	Yes	405 (I), 306 (II), 28 (I or II)
U.S. History: 1877 to Present (I) & *Colonization through Reconstruction* (II) [ISBN: 978-01-3331327-7, 978-0-13-330697-2]	James West Davidson, Michael B. Stoff; Emma J. Lapsansky-Werner, Peter B. Levy, Randy Roberts, Alan Taylor	Pearson	2016	Yes	163 (I), 158 (II)
The American Pageant, Volumes I & II, 14th ed. [ISBNs: 978-0-547-16659-9, 978-0-547-16658-2]	David M. Kennedy, Lizabeth Cohen, Thomas A. Bailey	Cengage	2010	No	145
America's History for the AP Course, 8th ed. [ISBN: 978-1457673825]	James A. Henretta, Eric Hinderaker, Rebecca Edwards, Robert O. Self	Bedford/St. Martin's	2014	No	82
Give Me Liberty! An American History, Volumes I & II, 3rd ed. [ISBNs: 978-0-393-11911-4, 978-0-393-11889-6]	Eric Foner	W.W. Norton & Co.	2011	No	77
Mastering the Grade 8 Social Studies TEKS (I) & *Mastering the TEKS in United States History Since 1877* [ISBNs: 978-1-935-02215-2, 978-1-935-02211-4]	Mark Jarrett, Stuart Zimmer, James Zilloran	Jarrett	2012	No	57 (I), 41 (II)
American History: Connecting With the Past, 15th ed. [ISBN: 978-0073513294]	Alan Brinkley	McGraw-Hill	2015	No	43
America: A Narrative History, 10th ed. [ISBN: 9780393265934]	David E. Shi, George Brown Tindall	W.W. Norton & Co.	2016	No	40
By the People [ISBN: 978-0-20574309-4]	James Fraser	Pearson		No	38
History Alive! The United States Through Industrialism [ISBN: 978-1-58371-931-2]	Diane Hart	Teachers' Curriculum Institute	2011	No	25
America: Past and Present, Volumes I & II, 10th ed. [ISBNs: 978-0205905195, 978-0205905478]	Robert A. Divine, T. H. Breen, R. Hal Williams, Ariela J. Gross, H. W. Brands	Pearson	2013	No	18
The Unfinished Nation: A Concise History of the American People, 8th ed. [ISBN: 978-1259287121]	Alan Brinkley	McGraw-Hill	2016	No	11
Visions of America: A History of the United States, 2nd ed. [ISBN: 978-0205092666]	Jennifer D. Keene, Saul Cornell, Edward T. O'Donnell	Pearson	2013	No	11

Note. Roman numerals I and II are used to indicate the first and second volumes of the same series by the same publisher.

Appendix D

Preparing the Text Data

We tried several OCR (optical character recognition) software packages and found that ABBYY FineReader worked best by a large margin. By performing manual checks, we found that ABBYY was highly accurate (fewer than 1 misrecognized character per page for main body text on average). As for the postprocessing of the text, we moved titles and subtitles into separate lines if they were concatenated with the following paragraph. We removed lines with fewer than 10 characters and ones with fewer than 5 tokens. We separate tokens using Natural Language Toolkit (NLTK)'s word tokenizer. The final corpus includes 7.6 million tokens overall and 102 thousand unique tokens.

Coreference Resolution Evaluation

One author annotated five randomly sampled excerpts from each textbook, focusing on coreference among mentions of people. Each of these excerpts were three paragraphs long and had 319.14 tokens on average. We considered a mention labeled by the model as correctly aligned to a human-labeled reference mention if its span completely includes the span of the human-labeled mention. We evaluated the model using B^3, a mention-based metric that compares gold clusters of mentions against the model's output clusters (Bagga & Baldwin, 1998):

$$\text{Precision} = \sum_{i=1}^{N} \frac{1}{N} \frac{\#\ \textit{of correct elements in entity i's output cluster}}{\#\ \textit{of elements in entity i's output cluster}}$$

$$\text{Recall} = \sum_{i=1}^{N} \frac{1}{N} \frac{\#\ \textit{of correct elements in entity i's output cluster}}{\#\ \textit{of elements in entity i's output cluster}}$$

Precision and recall range from 0 to 1, and F1 is the harmonic mean of the two. The coreference model achieved a F1 score of .704 (precision = .835, recall = .618), averaged across all textbooks on mentions involving individuals or groups. The precision is high, which means the coreference links detected by the model tend to be valid, but the recall is lower, meaning that the system fails to link some of the pronouns, names, or other expressions, which is a known limitation of coreference models (Durrett & Klein, 2013).

Named Entity Recognition Evaluation

To understand how well the NER tagger handles the textbook genre, we performed a manual evaluation. One author tagged five randomly selected coreference-resolved passages from each textbook, labeling only PERSON entities. Our NER tagger on this sample achieved an F1 score of .735 (precision = .768, recall = .706) when it comes to detecting the exact entity span. This is lower than the F1 score of .856 reported by SpaCy when evaluated on its original OntoNotes 5.0 dataset on all entity types, suggesting that training a model on NER-annotated history textbooks may be useful in future work.

Lexicons

NRC's VAD lexicon has labels for 20,000 words, Rashkin et al. (2016) has sentiment labels for about 950 verbs, and Sap et al. (2017) has power labels for 1700 verbs and agency labels for 2000 verbs. VAD scores are already numerical in NRC's lexicon, but for power and agency, we had to map Sap et al. (2017)'s discrete labels to numbers. For agency, a noun has a score of 1, 0, or −1 when it is the subject of a verb with high, neutral, or low agency, respectively. For power, the noun, as subject or object of a labeled verb, has a score of 1 for high power, 0 for neutral power, and −1 for low power. Sentiment connotation frames range from −1.0 to 1.0 and indicate the writer *w*'s perspective of a noun that is the subject *s* or the object *o* of a labeled verb (Rashkin et al., 2016). So, for a given noun, we averaged power and sentiment values over cases where it is a subject and those where it is an object of labeled verbs.

Word Embeddings

Word2vec Parameters. We set the number of dimensions for the model to 100 and window size to 5 (both of which are standard settings for this model), but we also experiment with other parameters and find that our results are robust to variability in parameter settings. We create embeddings for both single tokens (unigrams) and pairs of adjacent tokens (bigrams), not counting stopwords (e.g., *men and women* would be considered a bigram, because *and* is a stopword).

Obtaining Cosine Similarity Between Word Embeddings via Bootstrapping. To perform bootstrapping, we first split the corpus (all books combined) into sentences, yielding $N \approx$ 385,000 sentences total. Then, we randomly sample N sentences with replacement from our set of all sentences and train a word2vec model on this sample. We repeat this sampling and training process 50 times, yielding 50 separate word2vec models, each trained on different perturbations of the original data. All embeddings are projected onto the same vector space so that the distances between them correspond to semantic similarity. We estimate the similarity of words *A* and *B* by taking the mean cosine similarity of *A* and *B* across the 50 models.

Topic Modeling

Document Size. We conduct topic modeling at the sentence level. Other options, like using books or paragraphs, were not appropriate. As for books, we only have 15 of them, which yields noisy topic estimates. Using paragraphs was not an option for two main reasons. First, the OCR is imprecise at marking paragraph boundaries, as texts on the side are often combined with paragraphs in the main body. Second, the variance across books in terms of average paragraph length across books is significantly higher than in terms of sentence length, which becomes a confound when performing analyses across books. We calculate the average paragraph length (in terms of token count) for each book. These values range from 27 to 139 ($M = 68$, $SD = 34$). The high variability in terms of paragraph length is due to structural differences across books as well as differences in the OCR's precision across books. Doing the same for sentence length, we find that the values are more similar—they range from 13 to 21 ($M = 17$, $SD = 3$), showing much less variability. To put our numbers for mean sentence length in perspective, sentences in Britannica articles and Wikipedia articles on national histories are 20 and 22 words long on average, respectively, based on recent numbers reported by Samoilenko et al. (2018).

Hyperparameter Settings. We leave the hyperparameters of the LDA model at their default initial value set by MALLET (alpha = 5, beta = .01, number of iterations = 1,000). Traditionally, these parameters have been kept fixed through training, but certain recent papers advocate for hyperparameter optimization (Wallach et al., 2009), which entails updating the prior distribution for topics (alpha) and for words (beta) to better fit the data. However, we find that by turning hyperparameter optimization on, our topics become less coherent, dominated by frequent words, such as *american, unit*, and *war*. This kind of overfitting, resulting from hyperparameter optimization, has been described in Tang et al.'s (2014) analyses as well. One standard way of addressing this problem would be to create a domain-specific stopword list that includes these highly frequent words and remove them, but we do not want to remove words like *war*, since they do carry relevant topic information. Thus, in order to avoid overfitting and encourage a diversity of lexical items and topics, we keep our hyperparameters fixed through training. This results in more coherent topics and relatively similar topic probabilities.

Appendix E

Human Coding of Nonnamed People Mentions

We carried out the traditional method of hiring human coders for analyzing the mentions of non-named people. We asked three undergraduate research assistants to code *By the People: A History of the United States* (Fraser, 2016) for 1 hour and estimate the percentage of nonnamed people terms above. The range of human estimates were as follows: Women 10% to 30% (computational: 36.17%), Black 5% to 20% (computational: 43.87%), Latinx 3% to 15% (computational: 3.71%), and White 50% to 80% (computational: 35.35%). Clearly, human estimates vary widely across individuals when given the same task and textbook. This comparison demonstrates that although interpretation by expert humans is the gold standard for extracting labels and meaning from small amounts of text, coding entire textbooks for estimating representation of social groups still poses challenges. It is difficult for coders to process large amounts of text efficiently (i.e., fatigue, distraction, and time constraints limit human capability) without sacrificing good interrater reliability or requiring copious amounts of time.

Appendix F

The topics associated with men are more difficult to interpret in relation to topics associated with other groups, as it is possible *man* or *men* is used in an ungendered manner, such as in historical quotes. Here, for each topic associated with *man/men*, we include a random sample of 20 sentences where the corresponding topic has a high probability (>.3). The sentences are uniformly sampled across all books. Topic 3, as we mention in the paper, consists predominantly of quotes. We do not treat quotes differently from the main text in our analyses, as we believe that the choice of quotes is a core part of a textbook's framing of people as well.

Topic 1: soldier, thousand, die, kill, hundr, death, year, day, men, fight

1. Dubbed Operation Desert Storm, it lasted only four days—the "hundred-hour war" (see Map 40.5).
2. As a generation of young men returned from the fighting of the war and took up young men's day-to-day lives in the civilian workforce, some worried that perhaps some had left some's masculinity
3. Of the estimated 1,600,000 people who died at Auschwitz, about 1,300,000 were Jews.
4. Over 1,500 soldiers were awarded the Medal of Honor for 1,500 soldiers' heroic actions during The war.
5. Sir Banastre Tarleton escaped, but 110 British soldiers were killed and more than 700 were taken prisoner.
6. In this short period, seasoned sergeants did seasoned sergeants' best to turn raw recruits into disciplined, battle-ready GIs.
7. That night, however, there was sullen silence as thousands of wounded and dying soldiers left on the

battlefield moaned and shrieked in agony amid the corpses of thousands of wounded and dying soldiers left on the 'battlefield's friends.

8. Not counting the hundreds of thousands of injured and crippled, the one millionth American had died in a motor vehicle accident by 1951—more than all those killed on all the battlefields of all the nation's wars to that date.
9. "I have returned many times to honor the valiant men who died serving me."
10. The young ladies of the town . . . had collected and were sitting in the stoops and at the windows to see the noble exhibition of a thousand half-starved and three-quarters naked soldiers pass in review before the noble exhibition of a thousand half-starved and three-quarters naked soldiers.
11. In this battle, however, an equal number of Japanese civilians either killed Japanese civilians (Japanese civilians had been told of mass rape and torture if taken prisoner) or were killed by Japanese soldiers if Japanese civilians tried to surrender.
12. Volunteer soldiers fought only for short periods of time and then returned home.
13. To suppress the mobs, Lincoln rushed in Union troops who had just fought at Gettysburg; they killed more than a hundred rioters.
14. Three days of rioting ensued in which thirty-four people were killed, twenty-five of thirty-four people Black.
15. Fast-moving machinery caused injuries and even deaths.
16. Harrison won, but a mere month after delivering the longest inaugural address ever (two hours), Harrison succumbed to pneumonia and died.
17. Chivington's troops opened fire, killing between 150 to 200 Cheyenne and Arapaho men, women, and children.
18. After a second day of fierce fighting, the Confederates retreated to Corinth, leaving the enemy forces battered and exhausted.
19. "It was either that or the atomic bomb, and I didn't hesitate a minute, and I've never lost any sleep over a minute since."
20. The British, with only one soldier wounded, marched on to Concord.

Topic 2: human, natur, man, person, thing, moral, reason, believ, good, individu

1. This man seemed obsessed with the preservation of public virtue.
2. As the Democratic New York Herald said, "We can now thrash Mexico into decency at our leisure".
3. Any law that uplifts human personality is just.
4. Their environment—for example, camouflage coloring for a moth—these characteristics, since they are genetically transmissible, become dominant in future generations.
5. . . . Our decision about energy will test the character of the American people.
6. Was the purpose of conservatism, one writer wondered, to create the "free man" or the "good man"?
7. Moreover, Kennedy's sense of caution and restraint, painful and frustrating as it was to African American activists, had proved to be well-founded.
8. Carter's emphasis on human rights led Carter to alter the U.S. relationship with a number of dictators.
9. Individuals should work for self-realization by resisting pressures to conform to society's expectations and responding instead to individuals' own instincts.
10. Inductive arguments help us make predictions and form hypotheses that we can test to see if inductive arguments are true.
11. How can our country look for aught but ignorance and vice, under the existing state of things?
12. Adams and Jefferson themselves displayed reasonable dignity, but Adams and Jefferson 'themselves' supporters showed no such restraint.
13. I need not caution you that a great deal depends upon your own proper attention to you and that you are careful of good Conduct during Harvest.
14. Emerson produced a significant body of poetry, but Emerson was most renowned for Emerson's essays and lectures.
15. Party spirit makes the worst of everything that opposes party spirit's folly.
16. A man of keen intelligence, Marshall had a carefully cultivated reputation for arrogance and a low tolerance for mediocrity.
17. Resurgent racism offered a convenient explanation for the alleged "failure" of Reconstruction.
18. They could not go to the segregated amusement parks advertised on television, and "living constantly on tiptoe stance, never quite knowing what to expect next" were the reasons, King explained, "why we find it difficult to wait."
19. Many Americans believed in Anglo-Saxon superiority—that Americans were a "superior -race" that should rule others.
20. Drawing on the work of John Locke, the English philosopher, they insisted that God had given they certain natural and inalienable rights.

Topic 3: man, hand, boy, thing, back, day, eye, told, cloth, dress

1. A bunch of loafers don't stop to consider that on the WPA are men and women who have traveled places and seen things, been educated and found a bunch of loafers' jobs folded up and nothing to replace a bunch of loafers with.
2. The sun came like gold through the trees and over the fields, and I felt like I was in heaven.
3. But now, tonight, the dead were risen, Earth was reinhabited, memory awoke, a million names were spoken: What was so-and-so doing tonight on Earth?
4. Most conflicts involved only a few warriors intent on stealing horses or "counting coup"—touching an enemy body with the hand or a special stick.
5. Workers grabbed the fire pails from a ledge above the tables and poured water on "Fire, but to no avail."
6. He had he wrapped up in an upside-down American flag, telling us that every star in an upside-down American flag represented a state stolen from the Indians.
7. I opened a paper to-day in which he [Webster] pounds on the old strings [of liberty] in a letter to the Washington Birthday feasters at New York.
8. We mean to make things over, we are tired of toil for naught, With but bare enough to live upon, and never an hour for thought;
9. Indians used virtually every part of a male bison Indians: meat for food; hides for clothing, shoes, bedding, and shelter; muscles and tendons for thread and bowstrings; intestines for containers; bones for tools; horns for eating utensils; hair for headdresses; and dung for fuel.
10. He sounded in my heart the first trumpet call of the new time that was to be.
11. One New Englander said the embargo was like "cutting one's throat to stop the nosebleed."
12. Grabbing fire buckets hanging by their front doors, colonists formed a double line from a fire to a river, pond, or well.
13. Lucky indeed was the aspiring office seeker who could boast of birth in a log cabin.
14. Empty pockets turned inside out were "Hoover flags."
15. God hath sifted a nation that god hath might send Choice Grain into this Wilderness.
16. Being awakened from a sound sleep by a fire alarm, the metaphor chosen by Thomas Jefferson, evoked the magnitude of the crisis.
17. The Eagle also carries a shield with red and white stripes and a blue field.
18. In the bottom image, a poor woman exclaims, "Oh Dear!"
19. Jo happened to suit Aunt March, who was lame and needed an active person to wait upon Jo's.
20. Rockefeller advised others that the key to success was "Put all your eggs in one basket and then watch one basket."

Topic 4: men, women, famili, children, young, work, woman, home, mother, husband

1. Many rejected the community's tradition of arranged marriages, insisting on choosing arranged marriages' own husbands and wives.
2. How do you think the Great Depression changed Americans' view of themselves? Consider the roles of men, women, and children in society in the family.
3. If particular care and attention is not paid to the Ladies we are and will not hold we bound by Laws in which we have no voice, or Representation/^
4. While Beecher upheld high standards in women's education, Beecher and many others argued that young women should be trained not for the workplace but in the domestic—arts—managing a kitchen, running a household, and nurturing the children.
5. Women's paid labor was making up for the declining earning power or the absence of men in American households.
6. Many women also carefully managed household budgets.
7. Other frolics included corn-husking bees for men and quilting bees for women.
8. Carla Rojas' mother returned home two years later, but Carla Rojas decided to remain.
9. Even so, many women also had the added burden of keeping many women's families together emotionally with many women's husbands out of work.
10. Aspiring young doctors served for a while as apprentices to older practitioners and were then turned loose on aspiring young doctors' "victims."
11. New England males lived not only to see New England males' own children reach adulthood but also to witness the birth of grandchildren.
12. Some families tried to teach some families' children to read and write at home, although the heavy burden of work in most agricultural households limited the time available for schooling.
13. physician Benjamin Rush argued that young women should ensure young women's husbands' "perseverance in the paths of rectitude" and called for loyal "republican mothers" who would instruct "loyal

republican mothers" who would instruct "their sons in the principles of liberty and 'government's sons in the principles of liberty and government."
14. This meant leaving family and friends, and jobs or school.
15. Women work longer and harder than most men.
16. In the 1740s, the Reverend Samuel Chandler of Andover, Massachusetts, was "much distressed for land for his children," seven of his children young boys.
17. In slavery, African American women's bodies had been the sexual property of White men.
18. Conflict undoubtedly, this was because these men and women were so benumbed by poverty that these men and women had little strength to protest.
19. Letters to distant husbands reflected how terribly letters to distant husbands' wives missed distant husbands and how these long separations were changing women's role in a society that had prided itself on male dominance and female fragility.
20. The war also enabled women to enter previously male-dominated professions such as teaching, civil service, and nursing.

Textbook Sources

Appleby, J., Brinkley, A., Broussard, A., McPherson, J., & Ritchie, D. (2016a). *United States history since 1877*. McGraw-Hill Education.

Appleby, J., Brinkley, A., Broussard, A., McPherson, J., & Ritchie, D. (2016b). *United States history to 1877*. McGraw-Hill Education.

Bower, B., & Lobdell, J. (2011). *History Alive! The United States through industrialism*. Teachers' Curriculum Institute.

Brinkley, A. (2015). *American history: Connecting with the past*. McGraw-Hill Education.

Brinkley, A., Giggie, J., & Huebner, A. (2016). *The unfinished nation: A concise history of the American people*. McGraw-Hill.

Davidson, J., & Stoff, M. (2016). *United States history: Colonization through reconstruction*. Pearson.

De la Teja, J. F., Danzer, G. A., Klor de Alva, J. J., Wilson, L. E., & Woloch, N. (2016). *The Americans: United States history since 1877*. Houghton Mifflin Harcourt.

Divine, R., Breen, T., Williams, H., Gross, A., & Brands, H. (2013). *America: Past and present*. Pearson.

Foner, E. (2013). *Give me liberty! An American history*. W. W. Norton.

Fraser, J. (2016). *By the people: A history of the United States*. Pearson.

Henretta, J., Hinderaker, E., Edwards, R., & Self, R. (2014). *America's history*. Bedford/St. Martins.

Jarrett, M., Zimmer, S., & Killoran, J. (2014a). *Mastering the grade 8 social studies TEKS*. Jarrett Publishing.

Jarrett, M., Zimmer, S., & Killoran, J. (2014b). *Mastering the TEKS in United States history since 1877*. Jarrett Publishing.

Keene, J., Cornell, S., & O'Donnell, E. (2013). *Visions of America: A history of the United States*. Pearson.

Kennedy, D., Cohen, L., & Bailey, T. (2010). *The American pageant: A history of the American people*. Wadsworth Cengage Learning.

Lapsansky-Werner, E., Levy, P., Roberts, R., & Taylor, A. (2016). *United States history: 1877 to the present*. Pearson.

White, D., & Deverell, W. (2016). *United States history: Early colonial period through reconstruction*. Houghton Mifflin Harcourt.

Authors' Note

The first two authors are equal contributors.

Acknowledgments

We would like to thank the following individuals for helpful conversations, feedback, and ideas: Noah Smith, Sebastian Munoz-Najar Galvez, Lily Fesler, Julia Lerch, Dallas Card, Ramón Antonio Martínez, Hannah D'Apice, Arya McCarthy, Bonnie Krejci, Mark Algee-Hewitt, AJ Alvero, Julia Perlmutter, Christine Wotipka, and Morgan Polikoff. We are grateful for the support of the Melvin and Joan Lane Stanford Graduate Fellowship (to D.D.) and NSF Graduate Research Fellowship Grant No. DGE 1752814 (to L.L.).

Notes

1. We thank an anonymous reviewer for making this point.
2. We use county level votes as opposed to precinct- or congressional districts-level ones because their geographic granularity matches that of school districts the closest. There are 254 counties in Texas, and it is straightforward to map them to the 1,227 school districts. Precincts are more granular than counties, but since there are a lot of them (two to eight per county) and their boundaries are different than those of school districts, it is much less trivial to map precincts to school districts. We did not consider using congressional districts because they are significantly more coarse than counties and school districts (there are only 36 congressional districts).
3. We find that the county-level percentage of students identifying as being part of an ethnic minority (American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander) has a very strong correlation with Democratic vote-shares (Pearson $r = .756$, $p < .001$).
4. A head of a phrase is the element that determines its syntactic function. For example, the head of *Spanish soldier* is the noun *soldier*, which is why we know that the whole phrase is a noun phrase.
5. Stemming involves reducing inflected or derived word forms to their stem or root (e.g. *books* and *booking* to *book; immigrate, immigration* and *immigrant* to *immigr*). Stems do not need to be actual words or word roots, what matters is that related words map to the same stem, so that they are treated the same by the algorithm or model that uses them.

References

Anderson, C. B., & Metzger, S. A. (2011). Slavery, the Civil War era, and African American representation in US history: An analysis of four states' academic standards. *Theory & Research in Social Education*, *39*(3), 393–415. https://doi.org/10.1080/00933104.2011.10473460

Antoniak, M., & Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, *6*, 107–119. https://doi.org/10.1162/tacl_a_00008

Apple, M. W. (1992). The text and cultural politics. *Educational Researcher*, *21*(7), 4–19. https://doi.org/10.3102/0013189X021007004

Apple, M. W., & Christian-Smith, L. K. (2017). The politics of the textbook. In M. Apple & L. Christian-Smith (Eds.), *The politics of the textbook* (pp. 1–21). Routledge. https://doi.org/10.4324/9781315021089-1

Aronson, B., & Laughter, J. (2016). The theory and practice of culturally relevant education: A synthesis of research across content areas. *Review of Educational Research*, *86*(1), 163–206. https://doi.org/10.3102/0034654315582066

Ash, E., Chen, D. L., & Ornaghi, A. (2020). *Stereotypes in high-stakes decisions: Evidence from US Circuit Courts*. National Bureau of Economic Research. https://users.nber.org/~dlchen/papers/Stereotypes_in_High_Stakes_Decisions.pdf

Bagga, A., & Baldwin, B. (1998, May). Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference* (Vol. 1, pp. 563–566).

Bamman, D., O'Connor, B., & Smith, N. A. (2013). Learning latent personas of film characters. In *Proceedings of the 51st annual meeting of the association for computational linguistics: Vol. 1. Long papers* (pp. 352–361). Association for Computational Linguistics. https://www.aclweb.org/anthology/P13-1035.pdf

Banks, J. A. (2001). Approaches to multicultural curriculum reform. In J. E. Banks & C. M. Banks (Eds.), *Multicultural education: Issues and perspectives* (4th ed., pp. 225–246). John Wiley. https://www.wcu.edu/WebFiles/PDFs/ApproachestoMulticulturalCurriculumReform.pdf

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(January), 993–1022.

Blumberg, R. L. (2007). *Gender bias in textbooks: A hidden obstacle on the road to gender equality in education*. UNESCO.

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. *Foundations and Trends® in Information Retrieval*, *11*(2–3), 143–296. https://doi.org/10.1561/1500000030

Bromley, P., Meyer, J. W., & Ramirez, F. O. (2011). The worldwide spread of environmental discourse in social studies, history, and civics textbooks, 1970–2008. *Comparative Education Review*, *55*(4), 517–545. https://doi.org/10.1086/660797

Brown, A. L., & Brown, K. D. (2010). Strange fruit indeed: Interrogating contemporary textbook representations of racial violence toward African Americans. *Teachers College Record*, *112*(1), 31–67.

Card, D., Gross, J., Boydstun, A., & Smith, N. A. (2016). Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1410–1420). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1148

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016). Topic modeling for evaluating students' reflective writing: A case study of pre-service teachers' journals. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 1–5). Association for Computing Machinery. https://doi.org/10.1145/2883851.2883951

Clark, K., & Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2256–2262). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1245

Collins, R. L. (2011). Content analysis of gender roles in media: Where are we now and where should we go? *Sex Roles*, *64*(3–4), 290–298. https://doi.org/10.1007/s11199-010-9929-5

Cornbleth, C. (2002). Images of America: What youth do know about the United States. *American Educational Research Journal*, *39*(2), 519–552. https://doi.org/10.3102/00028312039002519

Crossley, S. A., & Kyle, K. (2018). Analyzing spoken and written discourse: A role for natural language processing tools. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 567–594). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-59900-1_25

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S. A., Russell, D. R., Kyle, K., & Romer, U. (2017). Applying natural language processing tools to a student academic writing corpus: How large are disciplinary differences across science and engineering fields? *Journal of Writing Analytics*, *1*, 48–81.

Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A. (2014). Mining texts, learner productions and strategies with ReaderBench. In A. Peña-Ayala (Ed.), *Educational data mining* (pp. 345–377). Springer. https://doi.org/10.1007/978-3-319-02738-8_13

Dee, T. S., & Penner, E. K. (2017). The causal effects of cultural relevance: Evidence from an ethnic studies curriculum. *American Educational Research Journal*, *54*(1), 127–166. https://doi.org/10.3102/0002831216677002

Devlin, R. (2018). *A girl stands at the door: The generation of young women who desegregated America's schools*. Basic Books.

Dowell, N. M., Graesser, A. C., & Cai, Z. (2016). Language and discourse analysis with Coh-Metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics*, *3*(3), 72–95. https://doi.org/10.18608/jla.2016.33.5

Dowell, N. M., Nixon, T. M., & Graesser, A. C. (2019). Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior Research Methods*, *51*(3), 1007–1041. https://doi.org/10.3758/s13428-018-1102-z

Dozat, T., Qi, P., & Manning, C. D. (2017). Stanford's graph-based neural dependency parser at the conll 2017 shared task. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 20–30). Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-3002

Durrett, G., & Klein, D. (2013). Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 conference*

on empirical methods in natural language processing (pp. 1971–1982). Association for Computational Linguistics.

Fast, E., Vachovsky, T., & Bernstein, M. S. (2016). Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Tenth International AAAI Conference on Web and Social Media 2016*. https://hci.stanford.edu/publications/2016/ethan/gender.pdf

Fesler, L., Dee, T., Baker, R., & Evans, B. (2019). Text as data methods for education research. *Journal of Research on Educational Effectiveness*, *12*(4), 707–727. https://doi.org/10.1080/19345747.2019.1634168

Field, A., Bhat, G., & Tsvetkov, Y. (2019). Contextual affective analysis: A case study of people portrayals in online #MeToo stories. In *Proceedings of the international AAAI conference on web and social media* (Vol. 13, No. 01, pp. 158–169). https://www.aaai.org/ojs/index.php/ICWSM/article/view/3358

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in linguistic analysis* (pp. 1–32). Philological Society.

FitzGerald, F. (1980). *America revised: History schoolbooks in the twentieth century*. Random House.

Foster, S. J. (1999). The struggle for American identity: Treatment of ethnic groups in United States history textbooks. *History of Education*, *28*(3), 251–278. https://doi.org/10.1080/004676099284618

Fredriksen, B., & Brar, S. (2015). *Getting textbooks to every child in sub-Saharan Africa: Strategies for addressing the high cost and low availability problem*. World Bank. https://doi.org/10.1596/978-1-4648-0540-0

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the U S A*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Goldstein, D. (2020, January 12). Two states. Eight textbooks. Two American stories. *The New York Times*. https://www.nytimes.com/interactive/2020/01/12/us/texas-vs-california-history-textbooks.html

Gordon, J., & Van Durme, B. (2013). Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on automated knowledge base construction* (pp. 25–30). Association for Computing Machinery. https://doi.org/10.1145/2509558.2509563

Gordy, L., & Pritchard, A. M. (1995). Redirecting our voyage through history: A content analysis of social studies textbooks. *Urban Education*, *30*(2), 195–218. https://doi.org/10.1177/0042085995030002005

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *Elementary School Journal*, *115*(2), 210–229. https://doi.org/10.1086/678293

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

Greaney, V. (2006). Textbooks, respect for diversity, and social cohesion. In E. Roberts-Schweitzer (Ed.), *Promoting social cohesion through education: Case studies and tools for using textbooks and curricula* (pp. 47–69). World Bank.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Hines, M. (2016). Learning freedom: Education, elevation, and New York's African-American community, 1827–1829. *History of Education Quarterly*, *56*(4), 618–645. https://doi.org/10.1111/hoeq.12213

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing*.

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. In *Proceedings of the 54th annual meeting of the association for computational linguistics: Vol. 2. Short papers* (pp. 591–598). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2096

Hoyle, A., Wolf-Sonkin, L., Wallach, H., Augenstein, I., & Cotterell, R. (2019). Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1706–1716). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1167

Hutchins, R. D. (2011). Heroes and the renegotiation of national identity in American history textbooks: Representations of George Washington and Abraham Lincoln, 1982–2003. *Nations and Nationalism*, *17*(3), 649–668. https://doi.org/10.1111/j.1469-8129.2011.00488.x

Joseph, K., Wei, W., & Carley, K. M. (2017). Girls rule, boys drool: Extracting semantic and affective stereotypes from Twitter. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 1362–1374). Association for Computing Machinery. https://doi.org/10.1145/2998181.2998187

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Lachmann, R., & Mitchell, L. (2014). The changing face of war in textbooks: Depictions of World War II and Vietnam, 1970–2009. *Sociology of Education*, *87*(3), 188–203. https://doi.org/10.1177/0038040714537526

Lerch, J., Bromley, P., Ramirez, F. O., & Meyer, J. W. (2017). The rise of individual agency in conceptions of society: Textbooks worldwide, 1950–2011. *International Sociology*, *32*(1), 38–60. https://doi.org/10.1177/0268580916675525

Loewen, J. W. (2008). *Lies my teacher told me: Everything your American history textbook got wrong*. New Press.

Lugini, L., Litman, D., Godley, A., & Olshefski, C. (2018). Annotating student talk in text-based classroom discussions. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications* (pp. 110–116). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0511

Mayfield, E., & Rosé, C. P. (2013). LightSIDE: Open source machine learning for text. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 146–157). Routledge.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with CohMetrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Mehrabi, N., Gowda, T., Morstatter, F., Peng, N., & Galstyan, A. (2019). *Man is to person as woman is to location: Measuring gender bias in named entity recognition*. arXiv. https://arxiv.org/abs/1910.10872

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 3111–3119). Neural Information Processing Systems Foundation.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th annual meeting of the association for computational linguistics: Vol. 1. Long papers* (pp. 174–184). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1017

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403. https://doi.org/10.1093/pan/mpn018

Moreau, J. (2010). *Schoolbook nation: Conflicts over American history textbooks from the Civil War to the present*. University of Michigan Press.

Morning, A. (2008). Reconstructing race in science and society: Biology textbooks, 1952–2002. *American Journal of Sociology*, *114*(Suppl. 1), S106–S137. https://doi.org/10.1086/592206

Munoz-Najar Galvez, S., Heiberger, R., & McFarland, D. (2019). Paradigm wars revisited: A cartography of graduate research in the field of education (1980–2010). *American Educational Research Journal*, *57*(2), 612–652. https://doi.org/10.3102/0002831219860511

National Center for Education Statistics. (n.d.). *Common Core of Data (CCD) universe files* (2019-052) [Data file]. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2019052

Nguyen, D. (2017). *Text as social and cultural data: A computational perspective on variation in text*. Universiteit Twente. https://doi.org/10.3990/1.9789036543002

Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., & Winters, J. (2019). *How we do things with words: Analyzing text as social and cultural data*. arXiv. https://arxiv.org/abs/1907.01468

Nicholls, J. (2003). Methods in school textbook research. *History Education Research Journal*, *3*(2), 11–26. https://doi.org/10.18546/HERJ.03.2.02

O'Connor, B., Bamman, D., & Smith, N. A. (2011). Computational text analysis for social science: Model assumptions and complexity. In *Second workshop on comptuational social science and the wisdom of crowds (NIPS 2011)*. https://homes.cs.washington.edu/~nasmith/papers/oconnor+bamman+smith.nips-ws11.pdf

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, Article 13. https://doi.org/10.3389/fdata.2019.00013

Ornaghi, A., Ash, E., & Chen, D. L. (2019). *Stereotypes in high stake decisions: Evidence from US Circuit Courts* (Working Paper 2). Center for Law & Economics.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning* (No. 47). University of Illinois Press.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Pingel, F. (2010). *UNESCO guidebook on textbook research and textbook revision*. UNESCO.

Porter, M. F. (2001). *Snowball: A language for stemming algorithms*. http://snowball.tartarus.org/texts/

Ramesh, A., Goldwasser, D., Huang, B., Daumé, H., III, & Getoor, L. (2014). Understanding MOOC discussion forums using seeded LDA. In *Proceedings of the ninth workshop on innovative use of NLP for building educational applications* (pp. 28–33). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1804

Rashkin, H., Singh, S., & Choi, Y. (2016). Connotation frames: A data-driven investigation. In *Proceedings of the 54th annual meeting of the association for computational linguistics: Vol. 1. Long papers* (pp. 311–321). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1030

Read, A., & Bontoux, V. (2016). *Where have all the textbooks gone? The affordable and sustainable provision of learning and teaching materials in Sub-Saharan Africa*. The World Bank. https://openknowledge.worldbank.org/bitstream/handle/10986/22123/9781464805721.pdf?sequence=1

Reich, J., Tingley, D., Leder-Luis, J., Roberts, M. E., & Stewart, B. (2015). Computer-assisted reading and discovery for student generated text in massive open online courses. *Journal of Learning Analytics*, *2*(1), 156–184. https://doi.org/10.18608/jla.2015.21.8

Rockmore, E. B. (2015, October 21). How Texas teaches history. *The New York Times*. https://www.nytimes.com/2015/10/22/opinion/how-texas-teaches-history.html

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Samoilenko, A., Lemmerich, F., Zens, M., Jadidi, M., Génois, M., & Strohmaier, M. (2018). (Don't) Mention the war: A comparison of Wikipedia and Britannica articles on national histories. In *Proceedings of the 2018 world wide web conference* (pp. 843–852). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186132

Sap, M., Prasettio, M. C., Holtzman, A., Rashkin, H., & Choi, Y. (2017). Connotation frames of power and agency in modern films. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2329–2334). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1247

Sarvarzade, S., & Wotipka, C. M. (2017). The rise, removal, and return of women: Gender representations in primary-level textbooks in Afghanistan, 1980–2010. *Comparative*

Education, *53*(4), 578–599. https://doi.org/10.1080/03050068.2017.1348021

Schmidt, S. J. (2012). Am I a woman? The normalisation of woman in US History. *Gender and Education*, *24*(7), 707–724. https://doi.org/10.1080/09540253.2012.674491

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. MIT Press.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the international conference on machine learning* (pp. 190–198). JMLR.

Tetreault, M. K. T. (1986). Integrating women's history: The case of United States history high school textbooks. *History Teacher*, *19*(2), 211–262. https://doi.org/10.2307/493800

Texas Education Agency. (n.d.). *Instructional materials reports*. https://tea.texas.gov/Academics/Instructional_Materials/Instructional_Materials_Allotment/Instructional_Materials_Reports/

The New York Times. (2017, August 1). *Texas election results 2016*. https://www.nytimes.com/elections/2016/results/texas

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and education in twenty-eight countries: Civic knowledge and engagement at age fourteen*. IEA Secretariat.

Vytasek, J. M., Wise, A. F., & Woloshen, S. (2017). Topic models to support instructors in MOOC forums. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 610–611). Association for Computing Machinery. https://doi.org/10.1145/3027385.3029486

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media* (pp. 454–463). Association for the Advancement of Artificial Intelligence.

Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In *Advances in neural information processing systems* (pp. 1973–1981)

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in neural information processing systems* (pp. 3261–3275). Neural Information Processing Systems Foundation.

Webster, K., Recasens, M., Axelrod, V., & Baldridge, J. (2018). Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, *6*, 605–617. https://doi.org/10.1162/tacl_a_00240

Zinn, H. (1984). *The twentieth century: A people's history*. Harper & Row.

Authors

LI LUCY is a PhD student in the School of Information at the University of California, Berkeley. Her research uses natural language processing and data mining to shed light on social behavior and interactions.

DOROTTYA DEMSZKY is a PhD student in linguistics at Stanford University, advised by Dan Jurafsky. Her research focuses on developing natural language processing methods to study semantics and social phenomena mediated through language, with applications to the domain of education.

PATRICIA BROMLEY is an assistant professor of education and (by courtesy) sociology at Stanford University. Her research focuses on the rise and globalization of a culture emphasizing rational, scientific thinking and expansive forms of rights, and spans a range of fields including comparative education, the sociology of education, organization theory, and public administration and policy.

DAN JURAFSKY is professor and chair of linguistics and professor of computer science at Stanford University. His research focuses on natural language processing as well as its application to the behavioral and social sciences.