Bridging human and machine scoring in experimental assessments of writing: tools, tips, and lessons learned from a field trial in education

Reagan Mozer
Bentley University

Luke Miratrixy
Harvard University

Jackie Eunjung Relyea
North Carolina State University

James S. Kim
Harvard University

In a randomized trial that collects text as an outcome, traditional approaches for assessing treatment impact require that each document first be manually coded for constructs of interest by human raters. An impact analysis can then be conducted to compare treatment and control groups, using the hand-coded scores as a measured outcome. This process is both time and labor-intensive, which creates a persistent barrier for large-scale assessments of text. Furthermore, enriching ones understanding of a found impact on text outcomes via secondary analyses can be difficult without additional scoring efforts. Machine-based text analytic and data mining tools offer one potential avenue to help facilitate research in this domain. For instance, we could augment a traditional impact analysis that examines a single human-coded outcome with a suite of automatically generated secondary outcomes. By analyzing impacts across a wide array of text-based features, we can then explore what an overall change signifies, in terms of how the text has evolved due to treatment. In this paper, we propose several different methods for supplementary analysis in this spirit. We then present a case study of using these methods to enrich an evaluation of a classroom intervention on young children's writing. We argue that our rich array of findings move us from "it worked" to "it worked because" by revealing how observed improvements in writing were likely due, in part, to the students having learned to marshal evidence and speak with more authority. Relying exclusively on human scoring, by contrast, is a lost opportunity.

VERSION: November 2021

Suggested citation: Mozer, Reagan, Luke Miratrix, Jackie Eunjung Relyea, and James S. Kim. (2021). Bridging human and machine scoring in experimental assessments of writing: tools, tips, and lessons learned from a field trial in education. (EdWorkingPaper: 21-493). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/ecs1-1n25

Bridging human and machine scoring in experimental assessments of writing: tools, tips, and lessons learned from a field trial in education

Reagan Mozer[1], Luke Miratrix[2], Jackie Eunjung Relyea[3], and James Kim[2]

[1]Bentley University

[2]Harvard University Graduate School of Education

[3]North Carolina State University

Author Note

Correspondence concerning this article should be addressed to Reagan Mozer, Department of Mathematical Sciences, Bentley University, 175 Forest St., Waltham, MA 02452. E-mail: rmozer@bentley.edu

Abstract

In a randomized trial that collects text as an outcome, traditional approaches for assessing treatment impact require that each document first be manually coded for constructs of interest by human raters. An impact analysis can then be conducted to compare treatment and control groups, using the hand-coded scores as a measured outcome. This process is both time and labor-intensive, which creates a persistent barrier for large-scale assessments of text. Furthermore, enriching ones understanding of a found impact on text outcomes via secondary analyses can be difficult without additional scoring efforts. Machine-based text analytic and data mining tools offer one potential avenue to help facilitate research in this domain. For instance, we could augment a traditional impact analysis that examines a single human-coded outcome with a suite of automatically generated secondary outcomes. By analyzing impacts across a wide array of text-based features, we can then explore what an overall change signifies, in terms of how the text has evolved due to treatment. In this paper, we propose several different methods for supplementary analysis in this spirit. We then present a case study of using these methods to enrich an evaluation of a classroom intervention on young children's writing. We argue that our rich array of findings move us from "it worked" to "it worked because" by revealing how observed improvements in writing were likely due, in part, to the students having learned to marshal evidence and speak with more authority. Relying exclusively on human scoring, by contrast, is a lost opportunity.

Keywords: text analysis, randomized controlled trial, automated scoring, argumentative writing

Bridging human and machine scoring in experimental assessments of writing: tools, tips, and lessons learned from a field trial in education

Introduction

Experimental research in education routinely relies on text collected from survey responses, written compositions, interviews, and other forms of discourse as a means to test psychological theories and to evaluate instructional practices. For example, recent studies have used written assessments, reflective journals, and dialogue transcribed from video recordings to investigate the influences of different learning interventions on students' scientific conceptions (Tsai and Chang, 2005; Hsu et al., 2011), critical thinking skills, and (Sharadgah, 2014), writing competencies (Fu et al., 2019; Jiang and Zhang, 2020). In order to make inferences about these types of cognitive and psychosocial abilities, the text documents produced in these settings must first be reduced to sets of statistical features that represent qualitative constructs relevant to the theory and/or intervention being assessed. This is typically done through a process of human scoring, whereby trained human raters apply a set of scoring rubrics to hand-code each document for the constructs of interest. This process, the current standard, is both time-consuming and limiting: even the largest human-coding efforts are typically constrained to measure only a small set of dimensions. Such efforts also represent a massive simplification of the data; written language encodes a rich set of information that captures far more than what can feasibly be extracted by a human rater.

Machine based text analytic and data mining tools offer one potential avenue to help facilitate research in this domain: we could, for example, supplement the "top-line" results of an impact analysis on human-coded outcomes with a secondary analysis of a suite of automatically generated auxiliary outcomes. To this end, modern methods based on natural language processing (NLP) allow for the automatic evaluation of an array of linguistic properties including measures of grammatical and mechanical accuracy, discourse structure, and lexical diversity. In addition to simple "surface features" of language, modern methods can just as easily compute measures of construct-relevant characteristics such as semantic meaning, coherence, and variations in prosody (i.e., patterns of rhythm and tone in language) (Yan et al., 2020). These features characterize a rich set of abilities reflected in student writing, many of which are "invisible" to a human rater coding for higher-level constructs (Pennebaker et al., 2014). By examining patterns of impacts across a wide array of such outcomes, we can explore how different micro-features of text (e.g., specific linguistic properties and/or psychological characteristics reflected in writing) may be driving top-line impacts.

In this spirit, this paper presents a comprehensive tutorial on automated scoring techniques (i.e., NLP, machine learning, text analysis) for experimental assessments of text. In particular, we consider how to leverage automated methods to expand the scope of what researchers might measure, in terms of a treatment impact, when using text as an outcome. This expansion can be either to supplement an existing analysis, as an aid to unpacking what may be driving an overall impact on a hand-coded outcome, or as an impact analysis in its own right, for those without the necessary resources for a full human-coding effort. We believe that by easing some of the burdens associated with coding qualitative constructs, we can radically enhance and expand the use of text data – a product that is tied to essential skills and that also can capture deep understanding of content – as an outcome in education evaluations.

To ground this idea, we present a case study based on a randomized controlled trial (RCT) recently conducted by Kim et al. (in press) to evaluate the Model of Reading Engagement (MORE), a content literacy intervention, on first and second graders' domain knowledge in science and social studies as reflected by their performance on an argumentative writing assessment. The researchers collected thousands of student-generated essays, which were then hand-coded by trained research assistants as a preliminary step to assessing treatment impact. Once all essays had been scored on a measure of holistic writing quality, they then estimated an average treatment effect (ATE) as the difference in these scores between students who received the MORE classroom intervention (i.e., treatment) and students who received the typical instruction (i.e., control). The estimated impacts on quality of students' argumentative writing essays are presented in Table 1 (further details of the data and analysis are given in a subsequent section). These findings, in effect size units indexed by average within-grade variation in outcomes, show robust and strong top-line impacts in both science and social studies. Notably, these impacts are most pronounced in the social studies domain, which has an estimated treatment effect of roughly double that of the science domain.

Overall, these results provide clear evidence that the MORE intervention improves the overall quality of students' argumentative writing essays. But through what mechanisms did this occur? Did treated students use more sophisticated vocabulary or more refined argument structures in their essays? Did they write with more confidence or sense of authority? The goals of the methods presented in this paper are to enrich such a top-line analysis by examining impacts on other aspects of the text as measured using automated methods. In particular, we ask:

1. Do students exposed to the intervention exhibit different underlying psychological states in their writing than those of the control group?
2. Do students use different words or phrases to convey their ideas in writing as a result of treatment?
3. With respect to both structure and content, are essays in the treatment group systematically more similar to "gold-standard" source texts than essays in the control group?
4. To what extent can automated scoring methods – specifically, a machine learning model trained to predict essay quality on prior data – recover the estimated impacts of the intervention on human-coded writing quality scores?

Overall, we find that by examining a variety of text-based features generated using automated methods, we can in fact uncover important potential mechanisms as to why the treatment was effective. These tools are general, and we seek to provide the details necessary so researchers can try them out on their own experiments. In that vein, we also provide a software package (link withheld for review) to make these tools accessible. We also gauge to what extent the results from an automated impact analyses taken in isolation, in this context, replicate the gold-standard findings based on careful human coding. This second exploration begins the conversation of how we might interpret impacts on machine-coded outcomes in general.

Traditional approaches for evaluating writing in education and psychology

Our case study, which examines a corpus of student-generated essays collected during a large-scale randomized trial, represents a broad class of experimental studies that rely on text as a basis for evaluating treatment impacts. A typical impact analysis in such settings might be interested in whether the texts observed in the treatment group are systematically different from those in control with respect to some qualitative construct (e.g. essay quality). In a standard (human-coded) approach, researchers would first develop a scoring rubric for the construct of interest that lists the criteria for coding a given document and outlines the characteristics of different score levels. This instrument would then be applied to a representative sample (or an entire collection) of texts by at least one human rater to generate a numerical outcome value for each document. Once all essays have been scored, an analyst could then estimate the ATE by calculating the difference in average scores between the treatment and control groups (possibly adjusting for demographic variables, other observed covariates, etc.). This process ultimately leads to an unbiased estimate of the average treatment effect with respect to the original construct of interest, assuming the human coding captured it successfully.

This approach could, in principle, be augmented or simply repeated to evaluate treatment impacts for any additional constructs (e.g., strength of argument, creativity, etc.). In practice, however, the time and resources required for coding can quickly make this approach untenable, and efforts are more often limited to scoring only one or a small set of outcomes. Herein lies a key philosophical issue. Text, by its nature, is complex and multifaceted; the inference that an intervention has led to meaningful changes in holistic judgments of text (e.g., essay quality) may reasonably be the result of changes along a number of different dimensions (e.g., grammar, vocabulary, organizational structure). In general, there are many factors that contribute to overall changes in writing, or more broadly, overall changes in the expression of ideas through language (Shermis and Burstein, 2003). Our aim is to leverage existing tools from the computational linguistics and NLP literatures to help unpack these complexities.

There is a long history of empirical research linking different, directly calculable measures of text to pertinent aspects of language acquisition, reading and writing proficiency, and content knowledge (Boyd and Pennebaker, 2015). Previous research in education has established connections between the linguistic properties of text and students' cognitive engagement (Joksimovic et al., 2014) as well as various learning outcomes (Crossley et al., 2016a). Other studies have found that more frequent use of specialized or specific terms (McNamara et al., 2010), longer words (Crossley et al., 2011), and more academic words (Douglas, 2013) are all indicative of higher quality writing. The expression of language also offers a window writers' social, cognitive, and affective states (Dowell and Graesser, 2014; Dowell et al., 2015). For instance, high rates of pronoun use have been associated with greater focus on one's self or one's social world, auxiliary verb use has been associated with a narrative language style, and the use of conjunctions has been associated with higher levels of cognitive complexity (Crossley, 2020). Similarly, Pennebaker and King (1999) found that a person's "linguistic style" (i.e., their habitual use of specific function words including pronouns, prepositions, articles, conjunctions, and auxiliary verbs in writing) serves as a reliable measure of individual differences. Function word use has also been identified as an important predictor of social status, culture, truthfulness, and depression (Chung and Pennebaker, 2007). In sum, there is clear evidence from the literature that suggests that the words writers use provide information about *how* they are thinking in addition to what they are thinking about (Pennebaker et al., 2014). Incorporating these measures as outcomes in an impact analysis may therefore reveal systematic differences between treatment and control groups on psychologically meaningful aspects of writing that would have otherwise been overlooked.

In the extreme, one might even imagine that we could use fully automated methods for impact assessments with text, using machine scored outcomes as our primary outcome of interest. In fact, a number of automated scoring systems have been successfully developed and deployed to address the cost of essay grading, particularly in the context of standardized assessments (e.g., Page, 1994; Burstein et al., 1998; Foltz et al., 1999; Attali and Burstein, 2006). While these approaches have serious risks, most particularly the risk of automatically coding constructs that do not have the depth of meaning or nuance that one might achieve by human scoring, automated methods also have a number of advantages. In addition to scalability, automated scoring tools offer more objectivity, consistency, and reproducibility than what can reasonably be achieved by human raters. Human scoring of complex constructs requires complex and nuanced judgment, which is subject to biases and inconsistencies that can quickly complicate large-scale assessments of text (Shermis and Burstein, 2003). By applying the same scoring algorithm across all text documents, automated methods therefore have the potential to reduce measurement error introduced by human raters (Correnti et al., 2020). These tools also present an opportunity to reduce the scale of human coding in assessments that aggregate scores across multiple human raters. Previous studies in the automated essay scoring literature have found that the level of agreement between human and machine generated scores is comparable to that achieved between two human raters (Shermis and Hamner, 2012; Rudner et al., 2006). For these reasons, there has been increasing debate in the research community about moving away from the exclusive reliance on human scoring as the gold standard, particularly in the context of educational assessments of writing (Correnti et al., 2020).

More on MORE, a Case Study from a Randomized Trial in Education

Before we describe our approaches for augmenting a classic impact evaluation of text data with automated, quantitative secondary analyses we present further details on our running example and case study of the MORE intervention.

RCT content literacy intervention

This study examines data from a cluster RCT that investigated the effectiveness of the intervention compared to typical instruction for promoting content literacy in science and social studies among first and second grade students. The intervention under investigation is designed to help young children acquire networks of related vocabulary words while they read and write about science and social studies content. Over a 20 lesson cycle, teachers use thematic lessons, concept mapping, and interactive read-alouds to enable their students to build networks of vocabulary knowledge and to transfer this knowledge during argumentative writing and collaborative research activities. In essence, the theory of change for the intervention is guided by the lexical quality hypothesis, which posits that learners must have deep knowledge of words (i.e., the word's spelling, pronunciations, meaning, and connections to related words) and efficiently access those words when they generating text for an argumentative writing task (Perfetti, 2007; Perfetti and Hart, 2002). Consistent with the lexical quality hypothesis, results from a large randomized controlled provided strong empirical evidence for the lexical quality hypothesis. In particular, the study findings showed positive treatment effects on vocabulary knowledge depth and argumentative writing in science and social studies (Kim et al., in press). Moreover, students' vocabulary knowledge depth mediated the intervention effects on argumentative writing outcomes. We extend these experimental findings by using machine-based text analytic tools to supplement the analysis of human-coded writing outcomes.

Data

Our initial sample was comprised of 5,494 first ($n = 2,787$) and second ($n = 2,707$) grade students from a total of 302 classrooms across 30 different elementary schools. As described in Kim et al. (2020), randomization was done using a within-school matched pairs approach; for each school, either the first grade classrooms or second grade classrooms were assigned to treatment. Within this sample, a total of 1,537 first graders and 1,349 second graders received the MORE classroom intervention (i.e., treatment), and the remaining 1,250 first graders and 1,358 second graders received the typical instruction (i.e., control). At the end of the study period, both groups completed an argumentative writing assessment on two topics – science and social studies – to evaluate students' knowledge of the elements and structure of an argument. The assessment consisted of a short source text to present background information relevant to the topic and an open-ended writing prompt. Both the source text and prompt varied by grade level and topic (see the Appendix for the prompts used in each assessment). For each topic, students were asked to respond to the prompt by making an argument and were reminded of the components of a good argument (it states your opinion, presents your reasons, explains your thinking, and has a conclusion).

At the end of the study period, students' hand-written responses to each essay prompt were transcribed to digital form and corrected for obvious spelling and punctuation errors by trained research assistants. Terms that were illegible or indecipherable were transcribed as XXX. As noted in Kim et al. (2020), this process has been shown to reduce presentation bias in human scoring stemming from poor handwriting skills (Graham et al., 2011) and is intended to help raters focus on the elements and structure of an argument in argumentative writing. For the primary impact analysis described in Kim et al. (in press), essays were then coded for the presence and/or strength of a variety of factors (e.g., statement of argument, use of evidence, conclusion). These component scores were then aggregated to produce a total writing score ranging from 0 to 7. Each essay was initially scored by two trained human raters. These scores showed a high degree of inter-rater agreement in both the science and social studies domains ($\kappa = 0.98$ and $\kappa = 0.97$, respectively). Prior to analysis, any discrepancies between raters were resolved by one of the authors to generate a final score for each essay. See Kim et al. (in press) for additional information about the study design and data collection process.

For the present study, we focus on evaluating the impacts of the intervention on student writing using data from the 2,929 students whose essay responses were scored by human coders. We excluded two students from the human-coded sample whose transcribed essay texts contained no words or characters. Within the sample, we observed a total of 2,746 science essays and 2,548 social studies essays. Each essay contained an average of 33.7 words (SD=21.3) and 177.7 characters (SD=111.2).

Preliminary impact analysis

We estimate top-level treatment impacts on students' writing with a hierarchical linear model with fixed effects for each school using cluster robust standard errors, clustered at the teacher (classroom) level. We fit separate models for writing outcomes in science and social studies, regressing the human-coded essay quality scores in each domain on an indicator for treatment and other observed covariates. In particular, for student i with teacher t in grade g of school j we have

$$Y_{itgj} = \alpha_j + \beta G_{itgj} + \tau Z_{gj} + \gamma X_{itgj} + \epsilon_{igtj},$$

where α_j are the school fixed effects, β is the difference in average quality scores between first and second grades, G_{itgj} is an indicator for grade two, τ is our parameter of interest (the average impact of treatment), Z_{gj} is an indicator for treatment assignment, γ is the effect of baseline pre-test (MAP/RIT) scores, denoted by X_{itgj}, and ϵ_{itgj} is our student level within-classroom residual. To investigate whether treatment was differentially effective by grade, we also extended each model to include an interaction term between students' grade-level and treatment assignment. Likelihood ratio tests indicated that these terms did not lead to significant improvements over the main effects model for either science or social studies writing. Results are presented on Table 1.

Methods

Given an experimental (or quasi-experimental) sample where text data constitute the outcome of interest, researchers might ask either of two general questions. The first question, typically based in substantive theory and initial interest that motivated the intervention itself, is confirmatory: "did the intervention change a specific aspect of the text?" The second question is more exploratory in nature:"which aspects of the text, if any, did the intervention change?"

To make these questions explicit, consider a randomized trial with N subjects, indexed by $i = 1, \ldots, N$. Subject i is assigned treatment Z_i, which equals 1 for subjects assigned to treatment (here, the classroom intervention) and 0 for subjects assigned to control (here, typical instruction). For each subject i, we observe a variable-length text document T_i, which has been scored by human raters to generate a numerical outcome Y_i. Following the Neyman potential outcomes framework (Rubin, 1974; Holland, 1986; Splawa-Neyman et al., 1923/1990) let $T_i(1)$ be the text response subject i would write if assigned to treatment and $T_i(0)$ the response if assigned control. These $T_i(z)$ are complex and multifaceted, being the text as written. Now let $Y_i(1)$ and $Y_i(0)$ be the scores that would be assigned to each possible response in a human scoring effort. These *potential outcomes* are counterfactuals of each other: we can observe, for any individual, only one of the two possible responses (and therefore only one of the two possible scores), depending on the treatment they receive (see, e.g., Gerber and Green, 2012, for more on potential outcomes in the context of randomized experiments). To observe $Y_i(1)$, for example, we would treat subject i who would produce the text $T_i(1)$. We then score $T_i(1)$, obtaining our final outcome. For subject i, the causal effect of treatment is defined as $Y_i(1) - Y_i(0)$; this is how treatment has changed one individual with the respect to the specific construct being measured by human raters. To draw inferences about the effects of treatment across all subjects, an impact analysis might then be interested in estimating, for example, the Average Treatment Effect (ATE) defined by $\tau = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - Y_i(0)$.

We can similarly use this approach to estimate the causal impacts of treatment on a set of secondary text-based outcomes measured using automated scoring methods. Consider, for instance, an arbitrary procedure or calculation f that returns a numerical feature $f(T)$ for an observed text response T. These f could, for example, summarize a specific linguistic, psychological, or mechanical characteristic of the organic text (e.g., total word count, readability score, number of grammatical errors). In our potential outcomes notation, let $(\tilde{Y}_i(0), \tilde{Y}_i(1)) = (f(T_i(0)), f(T_i(1)))$ denote the set of text-based outcomes that would be observed for subject i under assignment to treatment and control, respectively, for a given f. Treatment effects can then be defined by contrasts of these potential outcomes, just as with the human-scored outcomes.

Given this setup, we next, in the subsections below, turn to a general workflow that can be used to address the general research questions delineated above.

Generating auxiliary outcomes from text

As a first step for unpacking treatment impact on additional aspects of the text that may not have been directly assessed in the human coding process, we generate and curate a rich set of numerical features that capture different descriptive characteristics of the text. We build these features from the raw text itself using a mix of simple statistical procedures (i.e., counting frequencies of words) and off-the-shelf software packages that generate sets of features designed to capture different aspects of text. Given the "unstructured" nature of text, there is essentially no limit to the number of features possible. In the paragraphs that follow, we describe several common tools that can be used to generate a wide array of possible features that span a range of aspects one might be interested in. Taken as a whole, our final sets of features comprise a rich representation of ways the treatment may have altered the subjects in terms of how they generated text.

Simple summary measures. The simplest summary measures of text are based on term frequencies, which count the total number of times a given word or special character appears in a text. Once tabulated, these counts can then be aggregated to calculate univariate summaries of text such as lexical diversity. For example, the type-token ratio (TTR), defined as the number of unique words in a text (i.e., types) divided by the overall number of words (i.e., tokens), reflects the breadth of vocabulary used in a given document and can provide an indication of text cohesion (Dowell et al., 2016). In the same manner, one can also calculate the term frequencies for a set of pre-determined words and phrases. For example, if part of an intervention focused on encouraging the use of more scientific language, we might identify a set of words targeted by the intervention itself and count the rates of appearance of these words. Counts for these keywords could be used as standalone features, or aggregated to cumulative usage rates. Simple word counts can also be used to identify common rhetorical relations used in writing. For instance, previous research in the field of discourse analysis suggests that words such as “perhaps” and “possibly” are common cues used by writers to express a belief while developing an argument (Burstein et al., 1998). Tools for computing term frequencies and simple text summaries are available in most modern computing programs; see, for example, the Natural Language Toolkit (NLTK; Bird et al., 2009) for functionality in Python and the `quanteda` package (Benoit et al., 2018) for functionality in `R`.

Natural language processing. NLP describes a class of tools based on algorithmically recognizing linguistic patterns within text. See Allen (1995), Martin and Jurafsky (2009), and Clark et al. (2013) for general references. Standard NLP toolkits can be used to extract an array of features that measure both the syntactic and lexical characteristics of text. Basic techniques in this domain can segment texts by sentences or paragraphs, allowing for the calculation of measures such as average sentence length. Some of these measures, while simple to conceptualize, can be trickier to code than anticipated. For example, number of sentences can be difficult due to ambiguities between punctuation used separating sentences and for other purposes; consider, for example, the preceding ";" or a period used for "Mr." vs. the end of a sentence (or "vs.", for that matter). Other common NLP techniques can parse texts to extract part-of-speech categories (e.g., nouns, verbs, adjectives), sentence structures (e.g., verb phrases, clauses), and named entities (Benjamin, 2012; Collins-Thompson, 2014). Ratios of syntactic structure types per sentence are also commonly computed as measures of syntactic variety (Burstein et al., 1998).

In addition, there are a number of canonical metrics that have been developed for automatic evaluation of text with respect to constructs such as reading grade level and discourse cohesion, many of which can be easily computed using freely available NLP tools such as Coh-Metrix (Graesser et al., 2004) and the Tool for the Automatic Analysis of Text Cohesion (TAACO; Crossley et al., 2016b). For instance, formulas such as the Flesch Reading Ease Score (FRES; Flesch, 1948) and the Flesch-Kincaid Readability Score (FKRS; Kincaid et al., 1975) provide an index for readability and comprehension difficulty based on the word and sentence lengths found in a text. Similar statistical devices have been developed for cohesion, which measures the presence or absence of explicit cues used to establish connections between ideas expressed in a text, and coherence, which represents the sense of meaning and organization derived by the reader (McNamara et al., 2014).

Validated dictionaries. Researchers have also developed more complex dictionaries with collections of terms and phrases that are believed to be indicative of certain psychological processes and affective states reflected in writing. For example, Linguistic Inquiry Word Count (LIWC; Pennebaker et al., 2001) – a popular and commercially available tool for performing sentiment analysis – measures a given text document against a collection of established dictionaries built by experts in psychology and linguistics to generate a 94-dimensional numerical summary of the text. This output first includes indices for several syntactic and grammatical attributes (i.e., rates of pronoun or punctuation use, as discussed above), and also generates several psychological dimensions (e.g., anger, insight, power) as captured by the usage of different sets of words). It finally provides four high-level summary variables measuring analytical thinking, clout, authenticity, and emotional tone (Pennebaker et al., 2015) based on a proprietary calculation based on word appearance rates and possibly other measured aspects of the text. For each psychological dimension, the output generated by LIWC is at root the percentage of words within a given text that reflect that dimension as represented by a canonical list of words. LIWC has been validated in the psychological literature (Tausczik and Pennebaker, 2010) and has been widely applied in education research (Sell and Farreras, 2017; Robinson et al., 2013).

Model-based representations. In addition to features that are directly measurable from the text, one might also be interested in evaluating latent constructs captured in language data. For instance, statistical topic models such as the Latent Dirichlet Allocation (LDA; Blei et al., 2003) and the Structural Topic Model (STM; Roberts et al., 2013) are frequently used to represent documents as a mixture of K latent topics, denoted $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})$, where θ_{ik} measures the proportion of terms in document i that can be attributed to topic k. Each topic k is itself characterized by probability vector, β_k, defined over all terms in the vocabulary; that being said, they are typically described by just those terms estimated as most likely (or by the relative differences of these estimated probabilities, a quantity called "lift"). The intuition here is that we could then use these topic proportions for each document as features in their own right, with possible impacts on each of those features; this in principle allows for investigating whether the treatment group is writing more or less about certain topics than the control.

Similar features can be calculated using neural network embeddings, such as Word2Vec (Mikolov et al., 2013), which have found to be quite useful in machine translation (e.g., Branting et al., 2019). These embeddings allow for projecting each document, using a function trained on alternate bodies of text, into a dense lower-dimensional space. This again provides comparable fixed-length feature representations from variable-length pieces of texts (Mikolov et al., 2013), just as with topic modeling, above. Because these embeddings implicitly contain representations of word similarity, this approach can be especially appealing for short excerpts of text such as single sentences or paragraphs, as related words will end up with similar representations.

While simple in principle, there are some complications in using these within a randomized experimental context. In particular, as the generated models are fit to the text itself, they can vary depending on the impact of treatment on that text. For example, the treatment itself could alter the set of topics identified in the topic model, which makes describing the impact on treatment more difficult; see Egami et al. (2017) for further discussion. Furthermore, interpreting impacts on the generated features may be more difficult than more directly measured quantities that have human interpretability. That being said, these tools offer a rich suite of additional features one might consider.

Features generated for the case study

For the case study described in the preceding section, we first generated a variety of common text statistics, including frequencies of key words and phrases ("concept words"). Using functionality from the `quanteda` package in R (Benoit et al., 2018), we then computed a set of lexical diversity and readability scores for each document including indices for the TTR, FRES, and FKRS. Additional measures of local and global text cohesion including indices for sentence-level overlap and connectives indices were similarly generated using TAACO version 2.0.4 (Crossley et al., 2019). Using LIWC, each essay was also scored on an additional 44 structural categories, 46 underlying psychological dimensions, and four summary measures of writing style (analytical thinking, authenticity, clout, and emotional tone). This process resulted in a set of nearly 120 unique text-based outcomes, which we then compared between treatment and control groups within each grade level and subject.

Evaluating treatment impact across multiple measures of text

Given a rich feature set derived from machine measures of text, we can estimate treatment impacts on each of these features using standard inferential tools (e.g., regression, *t*-tests, or randomization tests). However, given the broadly constructed and large feature set, it will typically not be clear a-priori which aspects of the text would be sensitive to treatment. Further, simultaneously comparing treatment and control groups across a large number of text-based outcomes opens the door to a massive multiple testing problem. We therefore apply multiple testing corrections to screen which outcomes were impacted from the large set of generated outcomes in a manner that preserves a nominal error rate. This screened list identifies aspects of the text significantly impacted by treatment, including features that may not have initially been considered in the human coding process.

In particular, we employ the Benjamini-Hochberg (BH; Benjamini and Hochberg, 1995) procedure for controlling the False Discovery Rate (FDR), which is the expected proportion of false identifications among a set of findings deemed statistically significant. Specifically, after all hypothesis tests of interest have been performed, we apply the BH procedure to the resulting set of p-values to adjust for the number of "significant" results. One reason we prefer this approach over more classical, conservative methods such as the Bonferroni correction is that the FDR approach is scalable; it's performance remains strong as the number of tests grows (Glickman et al., 2014). Thus it is not necessary to determine the number of total hypothesis tests that will be conducted before the data analysis is performed. With text, while we do not expect impacts on our sets of features to be sparse, we also do not expect to find impacts on all of our generated features; it is reasonable to expect groups of related features to all be moved, to some degree, by the treatment. We therefore seek methods that can distinguish positive findings in a context where the expectations is that many hypothesis may well be null; the FDR method is well-suited for this aim. Finally, the FDR method can also flexibly accommodate collections of hypothesis tests that are believed to be correlated, which will usually be the case with summary measures of text.

Impact estimation in the case study

For each of the four grade level (grades 1 and 2) by domain (science and social studies) groups in the RCT case study, we fit a separate linear regression model (with cluster robust standard errors) for each of the 117 text-based outcomes generated from students' essay responses using automated scoring methods. Following the top-line impact analysis described earlier, each text-based outcome was regressed on treatment assignment and pre-test (MAP/RIT) scores. We drop school fixed effects as we are analyzing within grade, and thus have a full cluster randomized trial for each of our four evaluations. For each grade by domain, we then adjust the full set of findings using the FDR correction to identify the set of textual features that have been most prominently impacted by treatment. We can then compare patterns of impacts across the grades to examine consistency of findings. We also generated word counts for two user-defined canonical lists of vocabulary (these "concept words" are vocabulary actively taught in the intervention, and vocabulary implicitly taught). We calculated rates of word use for each of these lists, and included impacts on these rates as a separate secondary analysis.

Discovering differential use of words and phrases

As previously discussed, the words people use contain powerful indicators of how they are thinking as well as what they are thinking about (Pennebaker et al., 2014). In experimental contexts, a natural question therefore asks, "do the words used by individuals in the treatment group differ systematically from those used in the control group?" For this more exploratory question, we use Concise Comparative Summarization (CCS; Jia et al., 2014), as implemented by Miratrix and Ackerman (2016). CCS regresses a treatment indicator on the dynamically generated set of all possible words and phrases and identifies those words and phrases that are most predictive of treatment. These identified words and phrases are then taken as a summary of what separates the two groups of documents. Both Jia et al. (2014) and Miratrix and Ackerman (2016) found that the identified parts of text were interpretable by human readers, and in fact meaningfully differentiated the compared sets of documents. CCS is also at root also a predictive model: recent work by Kuang (2017) compared CCS to alternative strategies for text-based prediction and found that CCS outperformed other methods in terms of predictive accuracy and stability of the identified predictive text features when applied to political texts.

CCS is a sparse regularized regression method. In its usual form, it is a version of the LASSO method originally described in Tibshirani (1996) implemented using a particular normalization. Using CCS, we regress the treatment indicator onto the full set of words and phrases found in the text. The regularization means that it "costs" to use any given feature (a feature here being any possible word or phrase), and thus only those features particularly associated with treatment will be preserved. Regularization prevents overfitting, overfitting being the counterpart to the multiple testing problem discussed above, and the sparseness ensures that only a small subset of the likely millions of possible candidate phrases are in the end selected as the final summary. This makes the output tractable for human interpretation.

There are two concerns with using sparse regression to identify words and phrases of interest. First, text is highly correlated: words and phrases tend to appear together. This means that while one selected set of phrases might meaningfully divide the treatment and control corpora, others may be nearly as effective. From a human interpretation point of view, either set could be potentially useful in understanding aspects of how one set of documents is different from the other. Second, with machine learning methods there are usually a variety of tuning parameters and design decisions that one needs to set and make. Especially because text is so correlated, even subtle differences in these choices can result in different final sets of identified phrases. To get a rich range of features we therefore refit CCS with different settings of the different tuning parameters and options, harvesting the phrases from each. We then count the number of times each phrase is selected by a model, and present summaries of the phrase occurrences across the models as a sensitivity check. Phrases that consistently appear are taken as the more robust findings.

Some of these tuning parameters actually offer an opportunity for investigation. In particular, CCS has a parameter that controls how rare vs. common phrases are differentially penalized; as we change this parameter, the relative cost of different words and phrases as a function of their overall frequency in the corpus changes. We can use different values of this parameter to explore a range of summaries beginning with those focused on how more commonly used phrases differentially appear and ending with how very specific snippets of text systematically appear in one group vs. another.

Application to the case study

We first divided our data into four groups determined by each grade and subject combination. For each group we used CCS to regress the treatment indicator onto the set of all words and phrases to determine which words and phrases appear more (or less) often in the treated students writing than the control. We tune CCS using a permutation approach following Miratrix and Ackerman (2016), but permuting at the school level to account for the school-level clustering of treatment. The resulting selected phrases in each group provide additional context about the ways in which treatment is impacting *what* students write.

The promise and perils of automation without human coding

In some cases, if resources are tight, we might wonder what we might conclude with just the impacts on the machine-generated features. In this spirit we explore an alternative way of assessing top-line results by first calculating how similar each essay is to a gold standard exemplar, and then estimating impact of treatment on these similarity scores.

More broadly, it is of substantive interest to know to what degree different numerical measures, which are straightforward to measure but likely be more difficult than hand-coded constructs to interpret, are related to qualitative assessment. In this vein we also investigate to what degree the collection of features discussed above can predict the human scoring, and how well a model trained on pilot data and applied to these data would replicate our gold-standard human-coded results.

Importantly, due to the randomized treatment assignment, either of these methods provide a valid estimate of impact; in other words, if there was a systematic difference between treatment and control on any sort of predicted scores, that would be evidence that the treatment caused change. What would *not* be guaranteed, however, is whether the scores generated using automated methods would be in alignment with the original conceptual construct. It is possible, for example, that the scores are giving a distorted measure of the targeted construct. Thus, while the impact itself would be valid, the interpretation of what the impact signifies could be more difficult.

Assessing impact on essay similarity

Our first approach for a machine coded top-line measure for impact analysis, rooted in information theory and statistical thermodynamics, is based on comparing the "informational value" of a collection of documents. The idea is to measure how different one document is from one or more reference documents. If our target outcome is, in effect, a measure of essay quality, we can compare each student essay to a range of "gold standard" essays, scoring how related each text is to this set. The more similar, the more we might believe the student essay is also high quality. While any individual assessment would, of course, be missing a lot, we can look at whether there was an overall impact of the treatment on how similar student essays are to these reference essays. While clearly inferior to actual human coding, perhaps this rough proxy measure, when coupled with the subsequent steps, could still prove informative, at least when viewed in aggregate.

Following Mozer et al. (2020), we define the "descriptive similarity" of two text documents by the cosine of the angle between their corresponding term frequency vectors.

This metric has been widely used for text matching in computer science and has been shown to be highly predictive of human judgments about the descriptive similarity between texts (Mozer et al., 2020). Using this measure we calculate, for each student generated essay, the cosine distance between that document, represented as a vector of word counts, and the corresponding "gold-standard" document (i.e., the passage referenced in the essay prompt). A similarity score equal to 1 indicates that the essay is identical to the source text and a similarity of 0 means the essay is orthogonal (i.e., completely unrelated) to the source text. We therefore look for evidence that the treated group has systematically higher (more similar) scores.

Predicting outcomes using prior data

As a second exploration we used coded pilot data to learn a predictive model based on the full set of features. We then predicted outcomes for all our data and estimated impacts, treating the predicted values as a top-line result. Impacts on these proxy outcomes could serve as a way of cheaply assessing whether there was a treatment impact before fully coding the data.

Under this approach, we can completely black box our machine learning model that predicts scores. We could even use off-the-shelf essay grading tools such as e-rater (Attali and Burstein, 2006) or the Intelligent Essay Assessor (IEA; Foltz et al., 1999). All we need is a systematic process that converts the raw text to a number that we can use as a feature. The key is if we determine the method for scoring our essays without reference to the target data, we are protected from any concerns of bias as the scoring is independent of the randomization. Of course, if our scoring has no connection to human meaning, then even if we find an impact we may not know what that signifies. For example, a method of simply taking the length of the essay as our "quality" could result in a found impact, but it would not inform us as to whether the impact was meaningful in terms of the real goals of the intervention.

Application to the case study

We conduct impact analyses using both similarity scores and predicted scores for the present case study. To compute similarity scores, we use the source texts coupled with each of the four writing prompts (one for each grade and subject combination). In particular, we test the hypothesis that students exposed to the content literacy intervention would structure their argumentative writing essays in a manner that more closely resembled the structure and vocabulary of the source material (i.e., the passage the student was asked about in the writing prompt) by calculating the cosine similarity between each essay and this source text (rather than a "gold standard" essay). These similarity scores provide an objective and holistic measure of students' writing, in terms of content and syntactic structure, which we then use as a stand-alone outcome for impact analysis. It should be noted, however, that any realized impacts on this outcome – or on any outcome defined by comparisons across documents – are context-specific. While a higher similarity score in the case study suggests a "better" essay, it is a relative measure of writing and is only meaningful for comparing essays generated in the same context.

For prediction, we used data from a prior study (Kim et al., 2020) that examined the impacts of the same content literacy intervention on the quality of science writing in the first grade. Using these data, we constructed a predictive model that predicted human-coded essay quality scores from a set of machine measures of text. These features included simple summaries (e.g., word count), nominal measures of sentiment calculated using LIWC (e.g., use of "cognitive" words), and others derived from natural language representations that have been trained on separate corpora (e.g., GloVe word vectors; Pennington et al., 2014); in other words, we use the very features discussed above to predict the original human score in the pilot data. We then fit over 20 candidate machine learning models using functionality from the `caret` package in `R` (Kuhn et al., 2008). Candidate models included variants of classical linear models (e.g., simple linear regression implemented with and without boosting), common non-linear and/or non-parametric regression models (e.g., kernel support vector machines, k-nearest neighbors, etc.), tree-based models (e.g., random forests estimated with and without regularization), and neural networks. These candidate models were finally aggregated into a single ensemble learner – a predictive model formed by a weighted combination of multiple sub-models (Zhang and Ma, 2012) – using functionality from the `caretEnsemble` package in `R` (Deane-Mayer and Knowles, 2019); this final ensemble model will, given a vector of text features for an essay, produce a predicted score for that essay. We then take these predicted scores as proxy measures of our human coded outcome; as our predictive model is fit on a separate dataset, the predictions for our primary study are simply summary measures of the text as any other.

For both of our outcomes, we then estimate impacts using our same cluster-robust regression approach as we did for the human coded scores. Again, these scores are a summary measure of the text, and the summarization is independent of treatment (in fact in both cases it can be specified at baseline). Thus, any found impact on these scores is rigorous, randomized trial evidence of the treatment impacting writing; the clear interpretation of this result is what would potentially be lost without human coding.

Results

We next illustrate our approach for performing comprehensive impact analyses in randomized trials with text-based outcomes, by expanding upon our previously described top-line treatment impacts on students' holistic writing quality scores. We first investigate how the intervention may have impacted several more isolated aspects of students' writing, including students' underlying psychological states, through a series of auxiliary analyses using machine measures of the essay texts. We then examine use of vocabulary, both in a planned mode (using specified vocabulary lists) as well as a more open ended mode of asking what differences in word and phrase there are between the treatment arms. We finally synthesize these findings to gain a deeper understanding of how the intervention impacts first and second grade students' approach to writing and domain knowledge in science and social studies.

Impacts on psycholinguistic properties

To assess the impacts of the intervention on the psycholinguistic properties of students' writing, we analyzed each of our four grade by subject groups separately. For each of the 117 text-based outcome measures, we first collected the point estimates and corresponding 95% marginal confidence intervals for the standardized differences in means between treatment and control groups. For the ease of the reader, we present results for only a subset of the original features tested. Figure 1 shows the impact estimates for each of the four groups with respect to ten common text statistics, including total word count, readability score, and four summary measures of higher-level thinking generated using LIWC. Figure 2 then shows the estimated effects for the set of other auxiliary outcome measures that were found to have significant treatment impacts in at least one of the four groups, so as to aid comparison of the pattern of results across subjects and grade levels.

These findings uncover several important potential mechanisms as to *how* the treatment was effective for improving students' writing abilities in each domain. In the second grade, for instance, we see that essays from the treatment group tend to score higher on the dimension of analytical thinking and lower on the dimension of clout compared to essays from the control group. Comparisons in this domain also reveal a number of significant differences between treatment and control groups with respect to students' underlying psychological states. For example, we find that that first graders in the treatment group scored significantly higher on the LIWC dimension measuring cognitive processes compared to those in the control group. We see evidence of treatment impacts on different psychological dimensions in the second grade, with essays from the treatment group scoring significantly lower on the indices for social processes and family dynamics compared to essays from the control group.

Impacts on various structural and psychological aspects of students' writing are even more pronounced in the social studies domain, particularly for the set of features included in our post-hoc comparisons. Across both grade levels, essays from the treatment group score significantly higher on the dimension for clout, suggesting that students who received the classroom intervention are writing from perspective of higher expertise and with more confidence, on average, than students who received typical instruction. Treated students also write with significantly lower levels of authenticity and emotional tone than students in control, which indicates a more formal and distanced form of discourse. In addition, we find significant differences between treatment and control groups at both grade levels on a number of dimensions related to students' drives, needs, and motives. We also see that, among treated students, first graders use significantly more female references and second graders use significantly more male references compared to first and second graders assigned to control. Given that the essay prompts for each grade level asked students to make an argument in favor of one of two historical figures – Amelia Earhart or Sally Ride for first graders, and Henry Ford or Leonardo DaVinci for second graders – this finding might suggest that the intervention is promoting a more structured writing style and a greater attention to detail than typical instruction.

Differential use of words and phrases

We have two investigations of word use. The first is an impact analysis comparing the rates of use for pre-determined sets of words deemed relevant to the intervention. The second is a discovery process identifying words and phrases used differently in each treatment arm.

Table 2 summarizes the total number of occurrences of specifically taught and untaught "concept words" within each subject and grade level. As a sensitivity check, we examine both total number of occurrences as well as proportion of essays with at least one occurrence. First, we see very different patterns of results across our four groups in terms of baseline usage of words. For example, 16% of the grade 1 social studies students were using the untaught words, while 0% of the grade 2 social studies students were. This is likely due to the different word lists consisting of more or less common words. In terms of impacts, we see positive impacts on all word lists that had substantial baseline prevalence. For example, grade 1 social studies saw a near doubling, from 16% to 31%, on the proportion of first graders in the treatment group that used at least one *untaught* concept word in their written responses. (The taught words were more rarely used, but we still see a modest estimated increase, from 1% to 3%.) Among second graders, we see a similar, albeit smaller, difference between students' use of *taught* concept words.

We next turn to asking what words and phrases, more broadly, the treated students used. For each subject and grade level, the terms and phrases identified as distinguishing between treatment and control text, across all possible phrases, are presented in Figure 3.

Our main finding here concerns the elevated use of the terms "should" and "I think" by students in the treatment group compared to the phrase "my opinion," which appears more commonly in the control group. These differences, which are consistent across both subjects and grade level, might suggest that treated students approached the argumentative writing task with a greater sense of agency than students in the control group. In social studies writing, we also see an elevated use of "celebrate," indicating a tighter attention to the essay prompt.

Impacts on descriptive similarity

Figure 4 shows, for each subject and grade level, the distribution of descriptive similarity scores calculated between each essay and its corresponding "gold-standard" reference text(s). The corresponding effect estimates on average descriptive similarity, controlling for students' pre-test (MAP/RIT) scores, are presented in Table 3.

Once again we see positive treatment impacts in both subjects, with significant differences for all but first grade science. Overall, these findings seem to support the hypothesis that the classroom intervention leads students to compose their arguments in a more structured manner that includes incorporating one or more pieces of evidence to support their claims. We also see small to moderate location shifts in the distributions of similarity scores for both subjects and grade levels. Further, the distributions of similarity scores in treatment and control groups appear similar in terms of shape for each of the four groups; thus, the location shift is suggestive of a consistent improvement in writing among students who received treatment.

Impacts on proxy measure of overall writing quality

We finally turn to assessing how results from a machine-coded top-line impact analysis would compare to those of a human scoring effort. Table 4 summarizes the estimated treatment impacts as well as the estimated effects of grade level and students' pretest (MAP/RIT) scores on measures of essay quality generated from each of these approaches. We estimate impacts for each of the four groups with two models, one for science and one for social studies, that each include a grade by treatment interaction term. Our proxy is predicted quality scores for each essay calculated by applying predictive ensemble of machine learners trained on a sample of human-coded writing samples collected in a separate study by Kim et al. (2020).

We generally see positive treatment impacts across both subjects; most of these impacts are significant. For first grade science the ML predicted quality impact of 0.22 is very close to the human coded impact of 0.20, which is sensible in that the pilot data used to fit the proxy model were in fact data only from this domain and grade. For the other three groups, the effect sizes of the proxy estimates are generally lower in magnitude compared to those estimated using the human-coded outcomes, although they all agree in sign and two agree in significance. We also see a reduced coefficient for pretest score, further suggesting a lack of complete alignment with the machine generated scores and human gold standard. The ML predictions are measuring aspects of the writing that were impacted by treatment, in general, but there is reason to believe they are not capturing the exact same aspects as were targeted by the human construct.

Overall, the pattern of results suggests that machine coding can be used to demonstrate impacts on measures likely related to a target measure of interest. On the other hand, treating an impact on a machine measure with some skepticism seems warranted. The machine measures were only loosely correlated with human-scored outcomes, with correlations within the four groups ranging from 0.36 to 0.51. In the case of second grade science, the lack of alignment completely erased the impact estimate, underscoring how a miscalibrated automatic scoring model can fail to capture a true impact. Given the full set of these findings, perhaps there are middle roads one might take here, such as assessing initial impacts using a proxy measure before committing to a full human coding effort.

Discussion

To our knowledge, this study represents one of the first attempts to apply machine-based text analytic and data mining tools to enrich an experimental assessment on young children's argumentative writing outcomes. In general, the answers to each of our four main research questions underscore the idea that machine-based scoring and analytic tools should supplement rather than supplant human-coded writing scores. Overall, we have argued that these methods go beyond the question, "did the intervention work?" to address "how did it work?"

For RQ 1, we find clear evidence of qualitative, rather than quantitative, shifts in writing. That is, we see consistent treatment impacts on young children's underlying psychological states rather than on the technical aspects of their writing (e.g., word count, TTR, and readability). For RQ 2, we find suggestive evidence that treated students were likely to use vocabulary words that were directly taught by teachers in grade 1 science and grade 2 social studies. For RQ 3, we see descriptive similarity scores that are higher for treatment than control, suggesting that treated students moved toward the word choice in "gold-standard" reference texts. Both findings are consistent with the learning quality hypothesis that intervention enabled students to acquire deep word knowledge and to use those words while producing texts that similar to reference texts. For RQ 4, we find that supervised machine learning models can be used to estimate top-line treatment impacts that align with results based on human coding for grades and domains where prior data is available, but that the size of estimated impacts can be attenuated if the ML is not well tuned to the given context. These results also highlight the importance of having good training data to improve machine based predictions. In particular, in our case we had trained our predictive algorithm on only first grade writing data (the only available pilot data) and it appears as if differences in the writing across our considered domains made these predictions less aligned to the human coded quality scores in the other domains considered.

All code, written using a new software package (name and link to be provided) designed to increase the ease of these investigations, is available along with an extensive tutorial on the use of this package. We hope these tools can offer an accessible entry-point for educators interested in exploring patterns in student-produced texts, for instance, to identify common themes in course evaluations (Sheard et al., 2003).

Data and code availability

Upon acceptance, the authors will provide all replication materials used to generate the tables and figures presented throughout this manuscript along with the code tutorial that shows how to implement the different methods and analysis techniques described using publicly-available tools.

References

Allen, J. (1995). *Natural language understanding.* Pearson.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., and Matsuo, A. (2018). quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Boyd, R. L. and Pennebaker, J. W. (2015). A way with words: Using language for psychological science in the modern era. *Consumer Psychology in a Social Media World*, pages 222–236.

Branting, K., Weiss, B., Brown, B., Pfeifer, C., Chakraborty, A., Ferro, L., Pfaff, M., and Yeh, A. (2019). Semi-supervised methods for explainable legal prediction. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 22–31. ACM.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., Nolan, J., Rock, D., and Wolff, S. (1998). Computer analysis of essay content for automated score prediction: A prototype automated scoring system for gmat analytical writing assessment essays. *ETS Research Report Series*, 1998(1):i–67.

Chung, C. and Pennebaker, J. W. (2007). The psychological functions of function words. *Social Communication*, 1:343–359.

Clark, A., Fox, C., and Lappin, S. (2013). *The handbook of computational linguistics and natural language processing*. John Wiley & Sons.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Correnti, R., Matsumura, L. C., Wang, E., Litman, D., Rahimi, Z., and Kisa, Z. (2020). Automated scoring of students' use of text evidence in writing. *Reading Research Quarterly*, 55(3):493–520.

Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3).

Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., and Baker, R. S. (2016a). Combining click-stream data with nlp tools to better understand mooc completion. In *Proceedings of the Sixth International Conference on Learning Analytics & knowledge*, pages 6–14.

Crossley, S. A., Kyle, K., and Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, 51(1):14–27.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016b). 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.

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., and McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, 28(3):282–311.

Deane-Mayer, Z. A. and Knowles, J. (2019). *caretEnsemble: ensembles of caret models*.

Douglas, S. R. (2013). The lexical breadth of undergraduate novice level writing competency. *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquée*, 16(1):152–170.

Dowell, N. M., Graesser, A. C., and 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.

Dowell, N. M., Skrypnyk, O., Joksimovic, S., Graesser, A. C., Dawson, S., Gaševic, D., Hennis, T. A., de Vries, P., and Kovanovic, V. (2015). Modeling learners' social centrality and performance through language and discourse. *International Educational Data Mining Society*.

Dowell, N. M. M. and Graesser, A. C. (2014). Modeling learners' cognitive, affective, and social processes through language and discourse. *Journal of Learning Analytics*, 1(3):183–186.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., and Stewart, B. M. (2017). How to make causal inferences using texts. *arXiv preprint*.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221.

Foltz, P. W., Laham, D., and Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944.

Fu, Q.-K., Lin, C.-J., Hwang, G.-J., and Zhang, L. (2019). Impacts of a mind mapping-based contextual gaming approach on efl students' writing performance, learning perceptions and generative uses in an english course. *Computers & Education*, 137:59–77.

Gerber, A. S. and Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.

Glickman, M. E., Rao, S. R., and Schultz, M. R. (2014). False discovery rate control is a recommended alternative to bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology*, 67(8):850–857.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Graham, S., Harris, K. R., and Hebert, M. (2011). It is more than just the message: Presentation effects in scoring writing. *Focus on Exceptional Children*, 44(4).

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hsu, C.-Y., Tsai, C.-C., and Liang, J.-C. (2011). Facilitating preschoolers' scientific knowledge construction via computer games regarding light and shadow: The effect of the prediction-observation-explanation (poe) strategy. *Journal of Science Education and Technology*, 20(5):482–493.

Jia, J., Miratrix, L., Yu, B., Gawalt, B., El Ghaoui, L., Barnesmoore, L., and Clavier, S. (2014). Concise comparative summaries (ccs) of large text corpora with a human experiment. *The Annals of Applied Statistics*, 8(1):499–529.

Jiang, D. and Zhang, L. J. (2020). Collaborating with 'familiar'strangers in mobile-assisted environments: The effect of socializing activities on learning efl writing. *Computers & Education*, 150:103841.

Joksimovic, S., Gasevic, D., Kovanovic, V., Adesope, O., and Hatala, M. (2014). Psychological characteristics in cognitive presence of communities of inquiry: A linguistic analysis of online discussions. *The Internet and Higher Education*, 22:1–10.

Kim, J. S., Burkhauser, M. A., Mesite, L. M., Asher, C. A., Relyea, J. E., Fitzgerald, J., and Elmore, J. (2020). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *Journal of Educational Psychology*.

Kim, J. S., Relyea, J. E., Burkhauser, M. A., Scherer, E., and Rich, P. (2021). Improving elementary grade students' science and social studies vocabulary knowledge depth, reading comprehension, and argumentative writing: a conceptual replication. *Educational Psychology Review*, pages 1–30.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Kuang, C. Y. (2017). *Predictive and Interpretable Text Machine Learning Models with Applications in Political Science*. PhD thesis, UC Berkeley.

Kuhn, M. et al. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26.

Martin, J. H. and Jurafsky, D. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall Upper Saddle River.

McNamara, D. S., Crossley, S. A., and McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, 27(1):57–86.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miratrix, L. and Ackerman, R. (2016). Conducting Sparse Feature Selection on Arbitrarily Long Phrases in Text Corpora with a Focus on Interpretability. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(6):435–460.

Mozer, R., Miratrix, L., Kaufman, A. R., and Anastasopoulos, L. J. (2020). Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis*, 28(4):445–468.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of Experimental Education*, 62(2):127–142.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report, University of Texas at Austin.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., and Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12):e115844.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Pennebaker, J. W. and King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11(4):357–383.

Perfetti, C. A. and Hart, L. (2002). The lexical quality hypothesis. *Precursors of Functional Literacy*, 11:67–86.

Roberts, M. E., Stewart, B. M., Tingley, D., and Airoldi, E. M. (2013). The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*, pages 1–20. Harrahs and Harveys, Lake Tahoe.

Robinson, R. L., Navea, R., and Ickes, W. (2013). Predicting final course performance from students' written self-introductions: A liwc analysis. *Journal of Language and Social Psychology*, 32(4):469–479.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rudner, L. M., Garcia, V., and Welch, C. (2006). An evaluation of intellimetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4).

Sell, J. and Farreras, I. G. (2017). Liwc-ing at a century of introductory college textbooks: Have the sentiments changed? *Procedia Computer Science*, 118:108–112.

Sharadgah, T. (2014). Developing critical thinking skills through writing in an internet-based environment. In *Society for Information Technology & Teacher Education*

International Conference, pages 2178–2185. Association for the Advancement of Computing in Education (AACE).

Sheard, J., Ceddia, J., Hurst, J., and Tuovinen, J. (2003). Inferring student learning behaviour from website interactions: A usage analysis. *Education and Information Technologies*, 8(3):245–266.

Shermis, M. D. and Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Shermis, M. D. and Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. In *Annual National Council on Measurement in Education Meeting*, pages 14–16.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tsai, C.-C. and Chang, C.-Y. (2005). Lasting effects of instruction guided by the conflict map: Experimental study of learning about the causes of the seasons. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 42(10):1089–1111.

Yan, D., Rupp, A. A., and Foltz, P. W. (2020). *Handbook of automated scoring: Theory into practice*. CRC Press.

Zhang, C. and Ma, Y. (2012). *Ensemble machine learning: methods and applications*. Springer.

Table 1

Estimated effects (in effect size units) of grade level, pretest scores (MAP/RIT), and treatment assignment on average (human-coded) writing quality scores in science and social studies.

	Science	Social Studies
(Intercept)	2.14^{***}	1.80^{***}
	(0.13)	(0.11)
Grade 2	-0.65^{***}	-0.27^{***}
	(0.04)	(0.05)
Treatment	0.25^{***}	0.44^{***}
	(0.04)	(0.05)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2

Use of taught and untaught concept words in treatment and control, grouped by grade-level and subject. For each set of terms, columns show cumulative frequency (total number occurrences) and prevalence (number of essays with at least one occurrence) rates across essays in each treatment group.

Grade	Subject	Type	Frequency			Prevalence		
			Treatment	Control	Diff.	Treatment	Control	Diff.
1	Science	Taught	98	48	50	79 (10.9%)	38 (5.9%)	41 (5.0%)**
		Untaught	3	5	-2	3 (0.4%)	5 (0.8%)	-2 (-0.4%)
	Social	Taught	21	10	11	17 (2.5%)	7 (1.1%)	10 (1.4%)
		Untaught	280	117	163	215 (31.8%)	98 (16.1%)	117 (15.7%)***
2	Science	Taught	144	102	42	106 (16.1%)	86 (11.8%)	20 (4.2%)
		Untaught	6	2	4	6 (0.9%)	2 (0.3%)	4 (0.6%)
	Social	Taught	124	79	45	107 (18.0%)	68 (10.2%)	39 (7.8%)***
		Untaught	1	0	1	1 (0.2%)	0 (0.0%)	1 (0.2%)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3

Estimated treatment effects (in effect size units) on average descriptive similarity between student-generated essays and gold-standard reference texts, controlling for pre-test (MAP/RIT) scores.

	Science	Social Studies
Grade 1	0.05	0.18^{*}
	(0.07)	(0.09)
Grade 2	0.20^{*}	0.27^{**}
	(0.08)	(0.09)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4

Estimated effects (in effect size units) of grade level, pretest scores (MAP/RIT), and treatment assignment on average machine learning (ML) predicted quality scores (right) compared to estimated estimated effects on average human-coded quality scores (left) for each grade level and subject.

	Science		Social Studies	
	Human-coded	ML predicted	Human-coded	ML predicted
Grade 1 Baseline	2.14^{***}	2.35^{***}	1.85^{***}	1.95^{***}
	(0.05)	(0.05)	(0.06)	(0.04)
Grade 2 Baseline	1.40^{***}	2.14^{***}	1.47^{***}	2.17^{***}
	(0.05)	(0.06)	(0.05)	(0.08)
Pretest Score	0.50^{***}	0.38^{***}	0.52^{***}	0.41^{***}
	(0.03)	(0.02)	(0.02)	(0.02)
Treatment (Grade 1)	0.20^{**}	0.22^{**}	0.35^{***}	0.22^{**}
	(0.08)	(0.08)	(0.08)	(0.08)
Treatment (Grade 2)	0.31^{***}	0.05	0.51^{***}	0.20^{*}
	(0.09)	(0.09)	(0.09)	(0.10)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

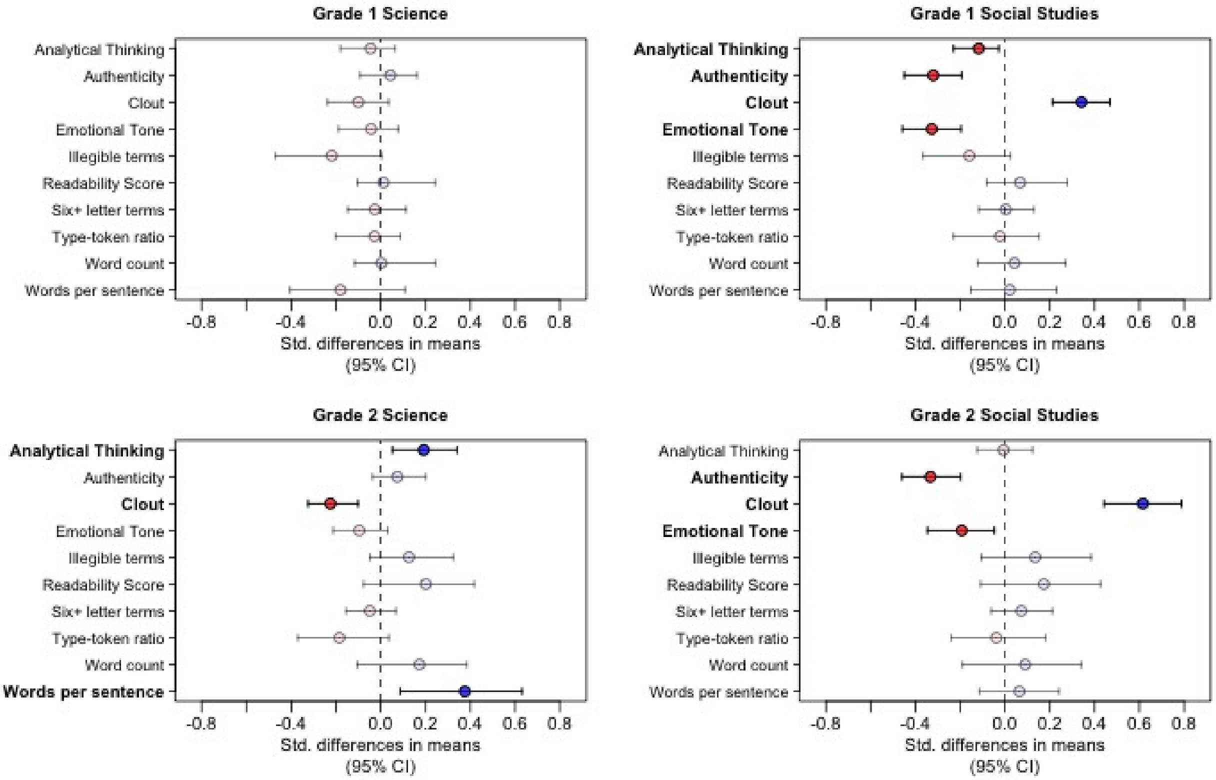


Figure 1

Standardized differences in means for a set of ten text-based outcomes capturing high-level properties of students' writing, grouped by grade and subject. For each group, bold feature names indicate statistical significance at the $\alpha = 0.05$ level (after corrections for multiple comparisons) and error bars show unadjusted *95% marginal confidence intervals. Red points indicate a negative estimated treatment impact and blue indicates positive estimated treatment impact.*

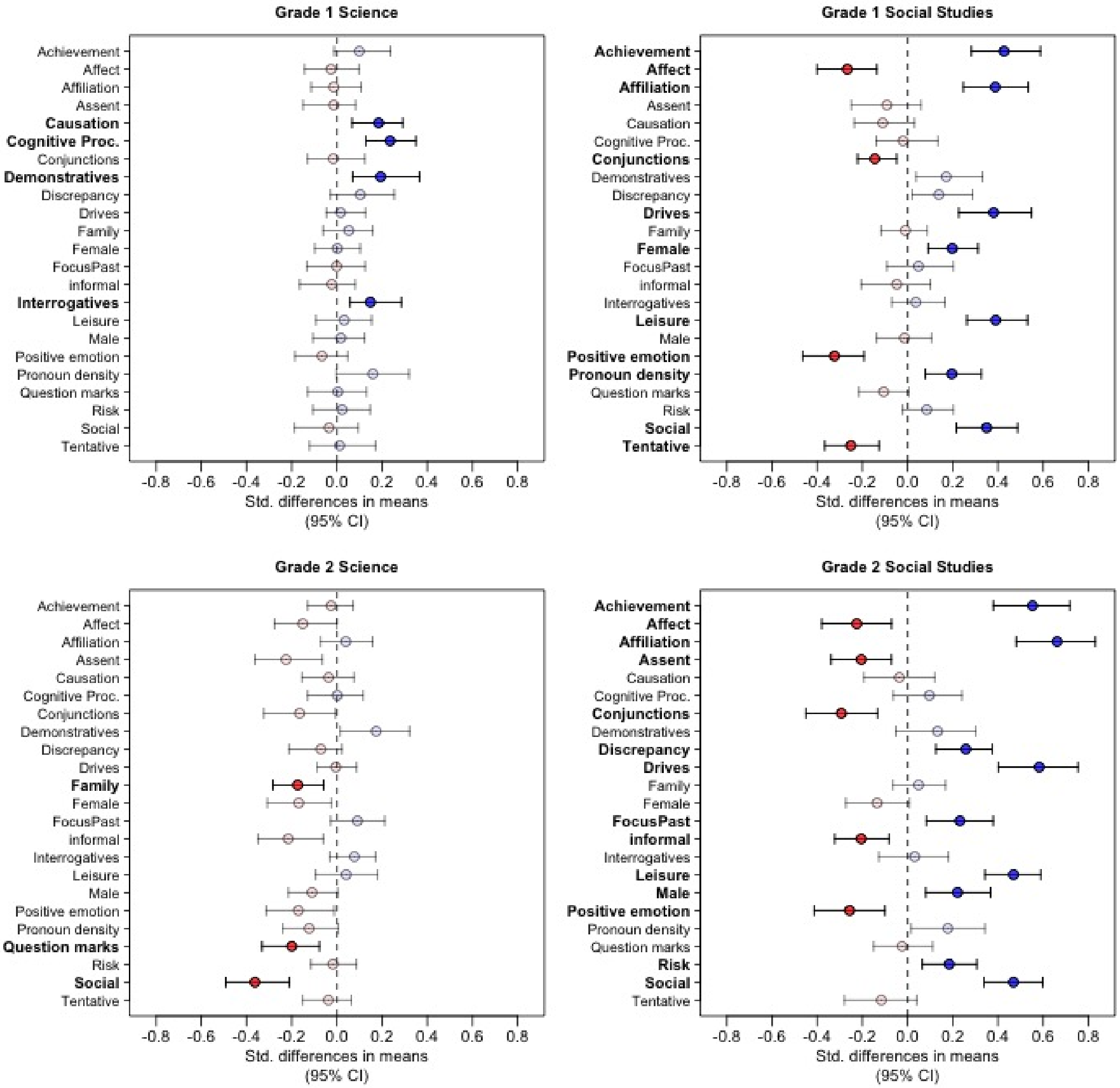


Figure 2

Standardized differences in means for psycholinguistic and structural features of students' essays found to be significant in at least one of the four grade by subject groups. For each group, bold feature names indicate statistical significance at the $\alpha = 0.05$ level (after corrections for multiple comparisons) and error bars show unadjusted *95% marginal confidence intervals. Red points indicate a negative estimated treatment impact and blue indicates positive estimated treatment impact.*

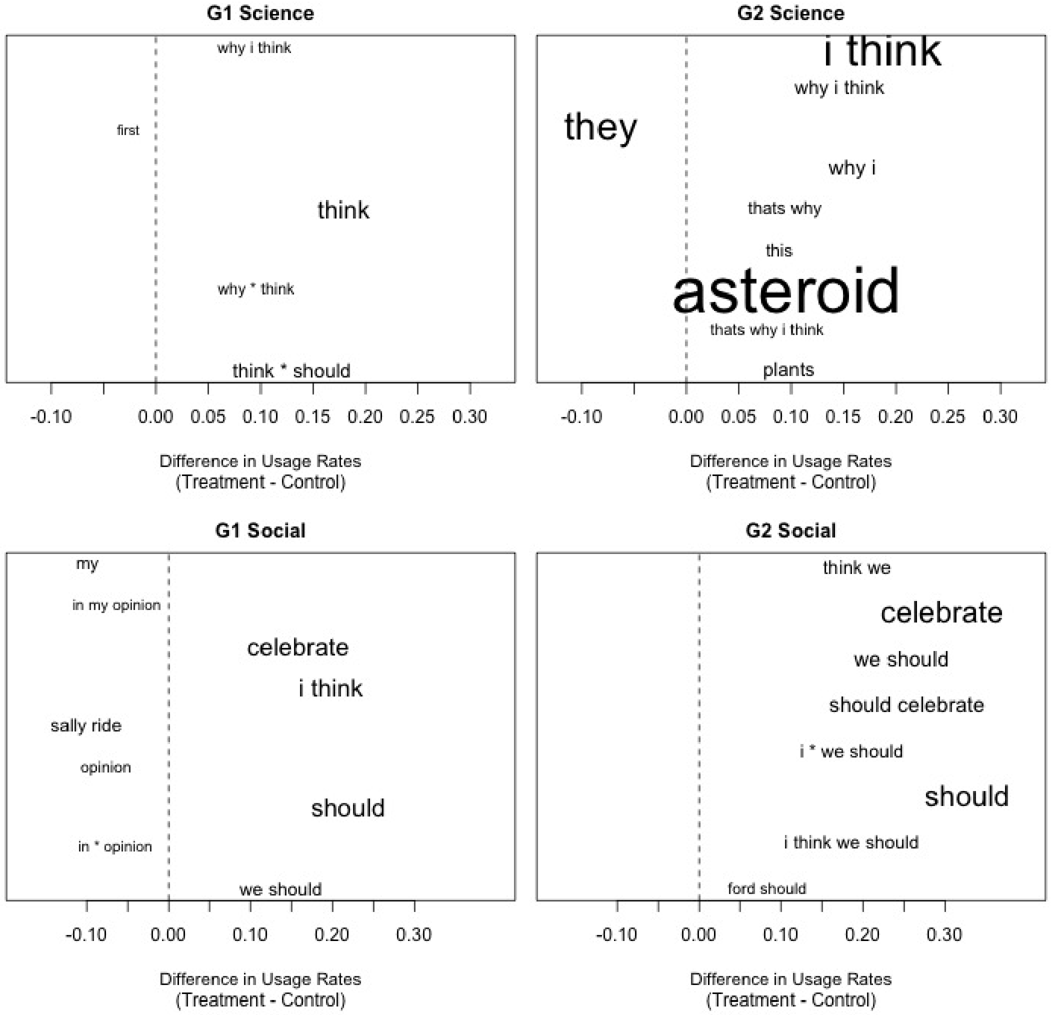


Figure 3

Terms and phrases identified as most predictive of assignment to treatment or control within each subject and grade level. Larger terms indicate greater cumulative frequency and terms that appear farther to the right indicate greater prevalence in the treatment group.

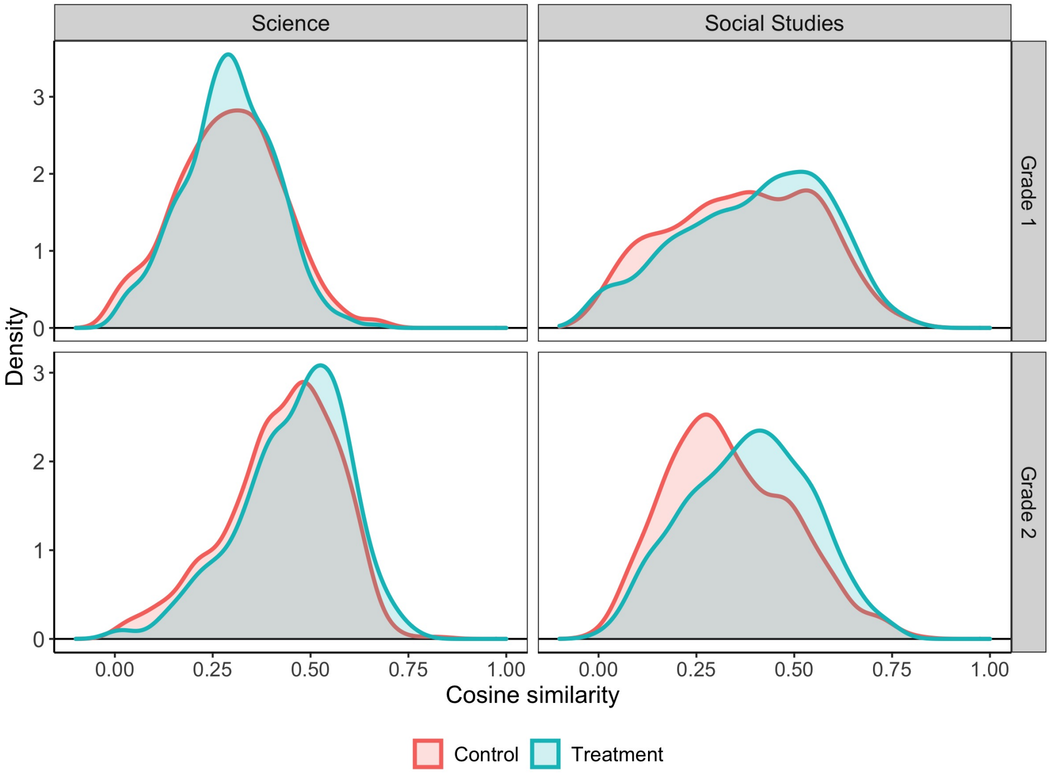


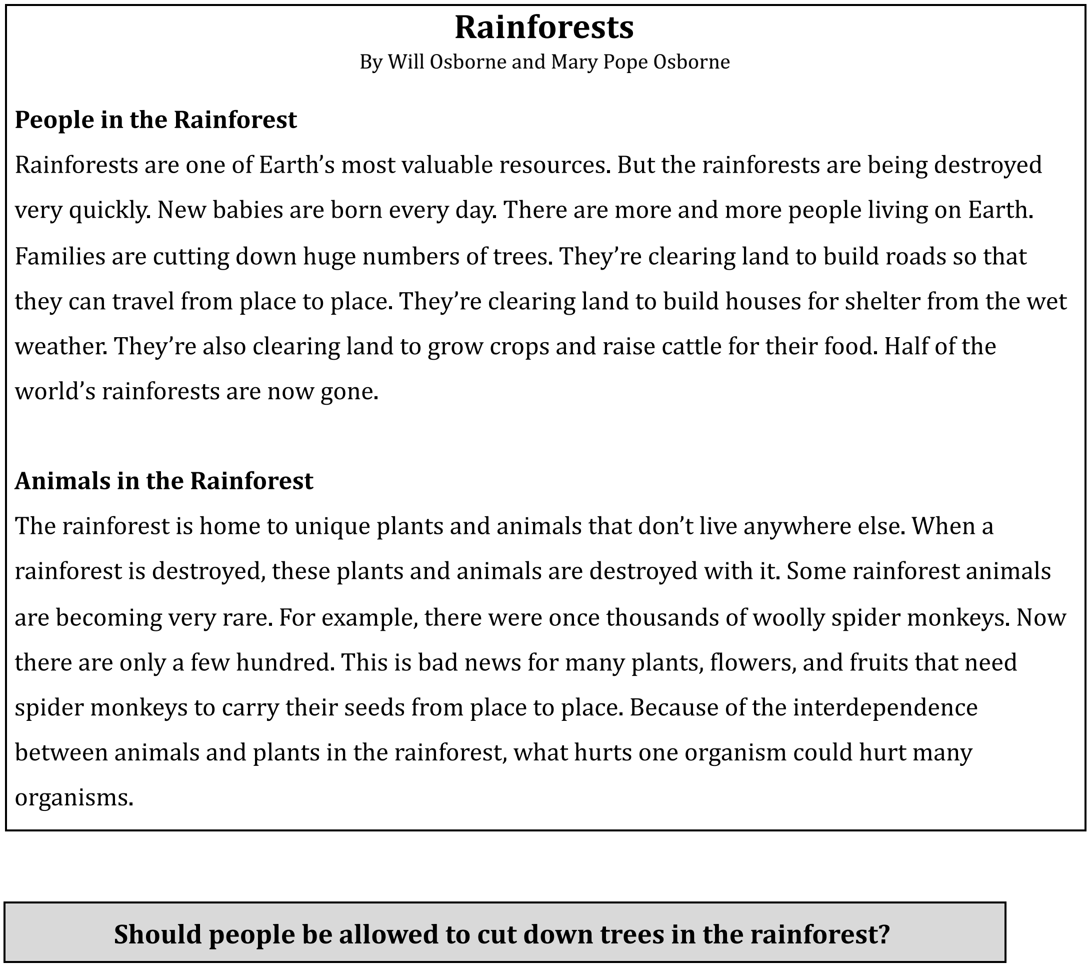
Figure 4

Kernel density estimates for the distribution of similarity scores in treatment and control for each subject and grade level. Higher values of cosine similarity suggest a higher degree of overlap with the source text.

Appendix

Essay prompts for each subject and grade level

For each of the argumentative writing assessments administered in the RCT study by Kim et al. (in press), students were asked to read a short passage describing a science or social studies topic and then respond to a writing prompt related to that topic. The passages and prompts for each subject and grade level are shown in Figures A1–A4.



Rainforests

By Will Osborne and Mary Pope Osborne

People in the Rainforest

Rainforests are one of Earth's most valuable resources. But the rainforests are being destroyed very quickly. New babies are born every day. There are more and more people living on Earth. Families are cutting down huge numbers of trees. They're clearing land to build roads so that they can travel from place to place. They're clearing land to build houses for shelter from the wet weather. They're also clearing land to grow crops and raise cattle for their food. Half of the world's rainforests are now gone.

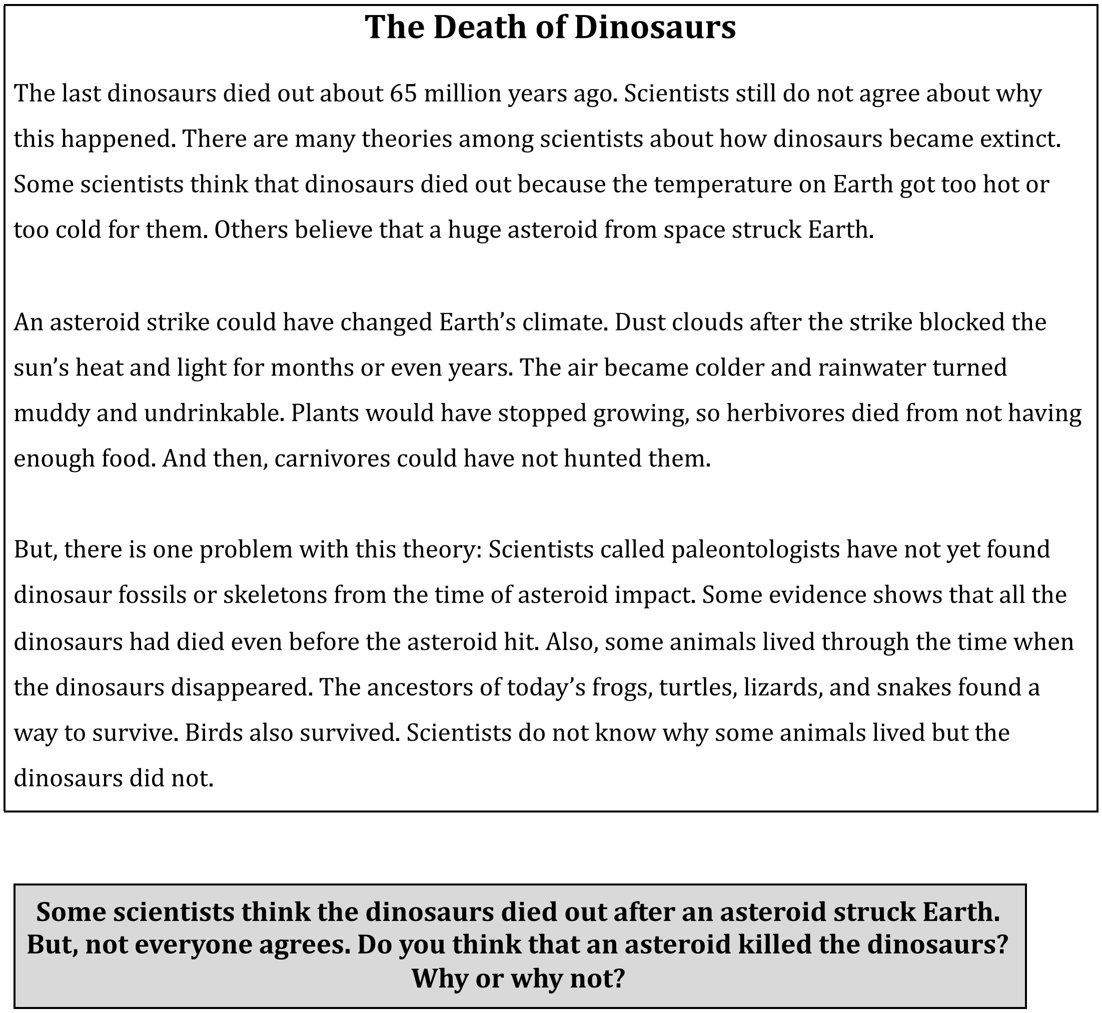
Animals in the Rainforest

The rainforest is home to unique plants and animals that don't live anywhere else. When a rainforest is destroyed, these plants and animals are destroyed with it. Some rainforest animals are becoming very rare. For example, there were once thousands of woolly spider monkeys. Now there are only a few hundred. This is bad news for many plants, flowers, and fruits that need spider monkeys to carry their seeds from place to place. Because of the interdependence between animals and plants in the rainforest, what hurts one organism could hurt many organisms.

Should people be allowed to cut down trees in the rainforest?

Figure A1

Passage and prompt for first grade science.



The Death of Dinosaurs

The last dinosaurs died out about 65 million years ago. Scientists still do not agree about why this happened. There are many theories among scientists about how dinosaurs became extinct. Some scientists think that dinosaurs died out because the temperature on Earth got too hot or too cold for them. Others believe that a huge asteroid from space struck Earth.

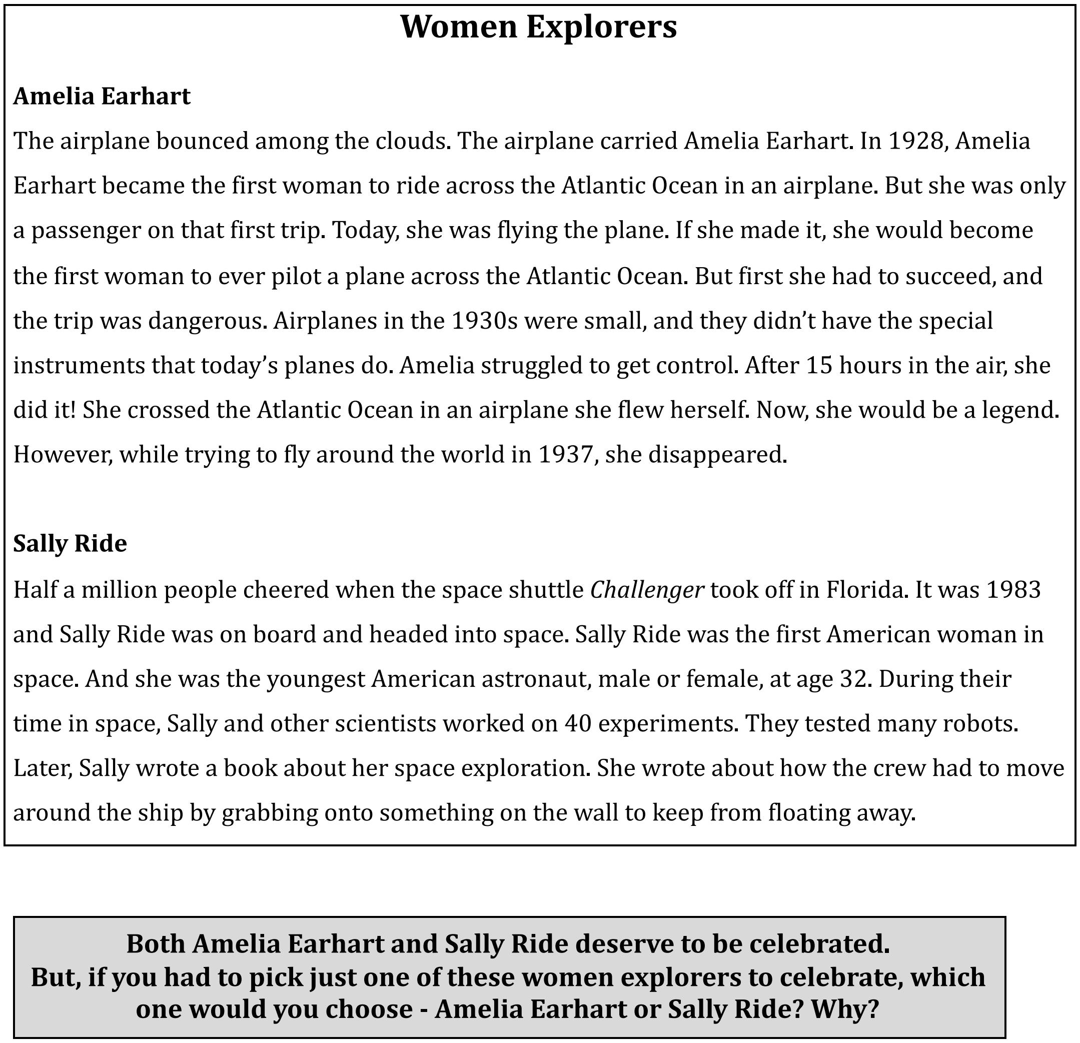
An asteroid strike could have changed Earth's climate. Dust clouds after the strike blocked the sun's heat and light for months or even years. The air became colder and rainwater turned muddy and undrinkable. Plants would have stopped growing, so herbivores died from not having enough food. And then, carnivores could have not hunted them.

But, there is one problem with this theory: Scientists called paleontologists have not yet found dinosaur fossils or skeletons from the time of asteroid impact. Some evidence shows that all the dinosaurs had died even before the asteroid hit. Also, some animals lived through the time when the dinosaurs disappeared. The ancestors of today's frogs, turtles, lizards, and snakes found a way to survive. Birds also survived. Scientists do not know why some animals lived but the dinosaurs did not.

Some scientists think the dinosaurs died out after an asteroid struck Earth. But, not everyone agrees. Do you think that an asteroid killed the dinosaurs? Why or why not?

Figure A2

Passage and prompt for second grade science.



Women Explorers

Amelia Earhart

The airplane bounced among the clouds. The airplane carried Amelia Earhart. In 1928, Amelia Earhart became the first woman to ride across the Atlantic Ocean in an airplane. But she was only a passenger on that first trip. Today, she was flying the plane. If she made it, she would become the first woman to ever pilot a plane across the Atlantic Ocean. But first she had to succeed, and the trip was dangerous. Airplanes in the 1930s were small, and they didn't have the special instruments that today's planes do. Amelia struggled to get control. After 15 hours in the air, she did it! She crossed the Atlantic Ocean in an airplane she flew herself. Now, she would be a legend. However, while trying to fly around the world in 1937, she disappeared.

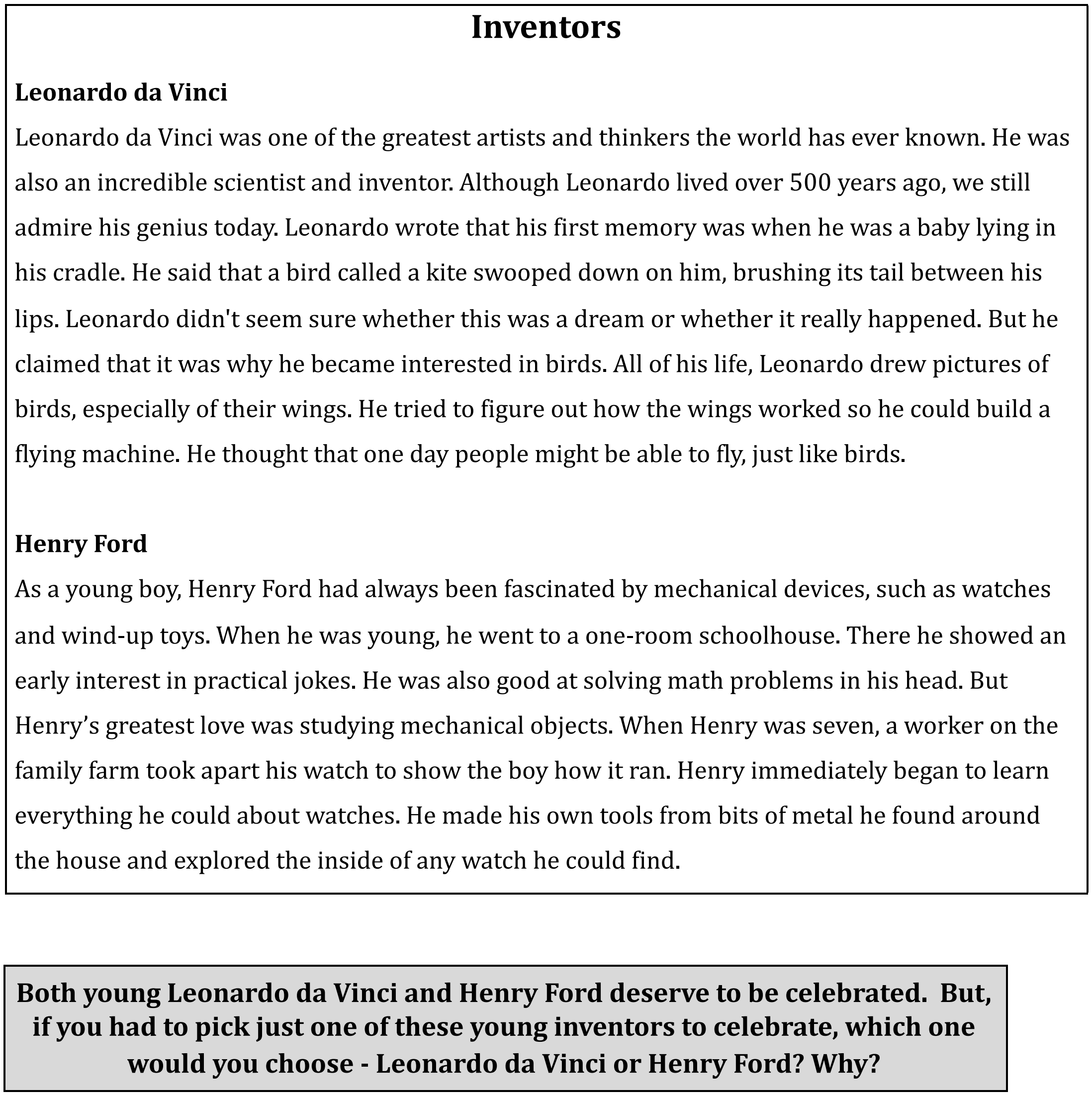
Sally Ride

Half a million people cheered when the space shuttle *Challenger* took off in Florida. It was 1983 and Sally Ride was on board and headed into space. Sally Ride was the first American woman in space. And she was the youngest American astronaut, male or female, at age 32. During their time in space, Sally and other scientists worked on 40 experiments. They tested many robots. Later, Sally wrote a book about her space exploration. She wrote about how the crew had to move around the ship by grabbing onto something on the wall to keep from floating away.

Both Amelia Earhart and Sally Ride deserve to be celebrated.
But, if you had to pick just one of these women explorers to celebrate, which one would you choose - Amelia Earhart or Sally Ride? Why?

Figure A3

Passage and prompt for first grade social studies.



Inventors

Leonardo da Vinci

Leonardo da Vinci was one of the greatest artists and thinkers the world has ever known. He was also an incredible scientist and inventor. Although Leonardo lived over 500 years ago, we still admire his genius today. Leonardo wrote that his first memory was when he was a baby lying in his cradle. He said that a bird called a kite swooped down on him, brushing its tail between his lips. Leonardo didn't seem sure whether this was a dream or whether it really happened. But he claimed that it was why he became interested in birds. All of his life, Leonardo drew pictures of birds, especially of their wings. He tried to figure out how the wings worked so he could build a flying machine. He thought that one day people might be able to fly, just like birds.

Henry Ford

As a young boy, Henry Ford had always been fascinated by mechanical devices, such as watches and wind-up toys. When he was young, he went to a one-room schoolhouse. There he showed an early interest in practical jokes. He was also good at solving math problems in his head. But Henry’s greatest love was studying mechanical objects. When Henry was seven, a worker on the family farm took apart his watch to show the boy how it ran. Henry immediately began to learn everything he could about watches. He made his own tools from bits of metal he found around the house and explored the inside of any watch he could find.

Both young Leonardo da Vinci and Henry Ford deserve to be celebrated. But, if you had to pick just one of these young inventors to celebrate, which one would you choose - Leonardo da Vinci or Henry Ford? Why?

Figure A4

Passage and prompt for second grade social studies.