Research
Report

Inside SourceFinder: Predicting the Acceptability Status of Candidate Reading-Comprehension Source Documents

Kathleen M. Sheehan

Irene Kostin

Yoko Futagi

Ramin Hemat

Daniel Zuckerman

Research &
Development

August 2006
RR-06-24

Inside SourceFinder:
Predicting the Acceptability Status of Candidate Reading-Comprehension Source Documents

Kathleen M. Sheehan, Irene Kostin, Yoko Futagi, Ramin Hemat, and Daniel Zuckerman
ETS, Princeton, NJ

August 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This paper describes the development, implementation, and evaluation of an automated system for predicting the acceptability status of candidate reading-comprehension stimuli extracted from a database of journal and magazine articles. The system uses a combination of classification and regression techniques to predict the probability that a given document will be deemed acceptable for use in completing a specified passage-creation assignment by at least one test developer. The text features that form the basis of the estimated models are automatically extracted by natural language processing techniques. Model performance is evaluated by comparing the proportion of acceptable documents located with the screening capability turned on to the proportion of acceptable documents located with the screening capability turned off. The evaluation suggests that the estimated models have succeeded in capturing useful information about the characteristics of texts that affect test developers' ratings of source acceptability and that they can help test developers find a greater number of high-quality sources in less time.

Key words: Content vector analyses, GRE®, reading-comprehension stimuli, source-acceptability modeling, SourceFinder

Table of Contents

List of Tables

List of Figures

Page

Introduction

For many common verbal item types, the item-writing process begins with a search for appropriate source material. For example, all of the reading passages on the verbal section of the Graduate Record Examinations® (GRE®) General Test are developed from previously published source texts extracted from books, scholarly journals, newspapers, or magazines. Experienced test developers report that the task of finding such source material is both difficult and time-consuming. For every acceptable source document located, a much larger number of unacceptable documents will have been retrieved, evaluated, and rejected (Passonneau, Hemat, Plante, & Sheehan, 2002). Experienced test developers also note that there is an inverse relationship between the acceptability status of a candidate source document and the time and effort needed to craft a set of items from that document. One test developer described this relationship as follows: "The single most important component of Verbal test development is the source. If you've identified a great source, then the rest of the writing and reviewing process seems to fall into place almost effortlessly; if you start with a mediocre source, then the problems with it will often follow you from the writing process all the way to the set's final review"A. Craig (personal communication, March 2, 2003).

An automated system for enhancing the efficiency of the source-selection process was developed at ETS in 1999 (Bauer & Jha, 1999; Jha, 2001). This system, called SourceFinder, employed a Web-crawling technique to locate potential source documents for use in developing reading-comprehension stimuli for the GRE General Test. Users initiated a search for needed source material by entering a start-up Web page. The system then downloaded potential source documents from that page and from other linked pages and sites. Retrieved documents were stored in a database for later consideration. Access to the database was provided by a graphical user interface (GUI) that also included an interactive document-screening capability. For example, a user could elect to consider only those documents with an average paragraph length (in words) above a specified threshold value, and with an average word length (in characters) above a second specified threshold value. A noncompensatory approach was used to implement these constraints. That is, a document was classified as being in compliance with the specified criteria only if *all* of the specified features fell above the specified threshold values. As is noted in Passonneau et al. (2002), this capability was designed to insure that all retrieved documents met minimum acceptability standards.

The SourceFinder system has undergone two major revisions since 1999. The first major revision was implemented in June 2003; the second was implemented in June 2005. This paper provides a description and evaluation of the design changes implemented in the June 2003 revision. A subsequent paper (Sheehan, Kostin, & Futagi, 2006) documents the additional modifications and enhancements implemented in the June 2005 revision.

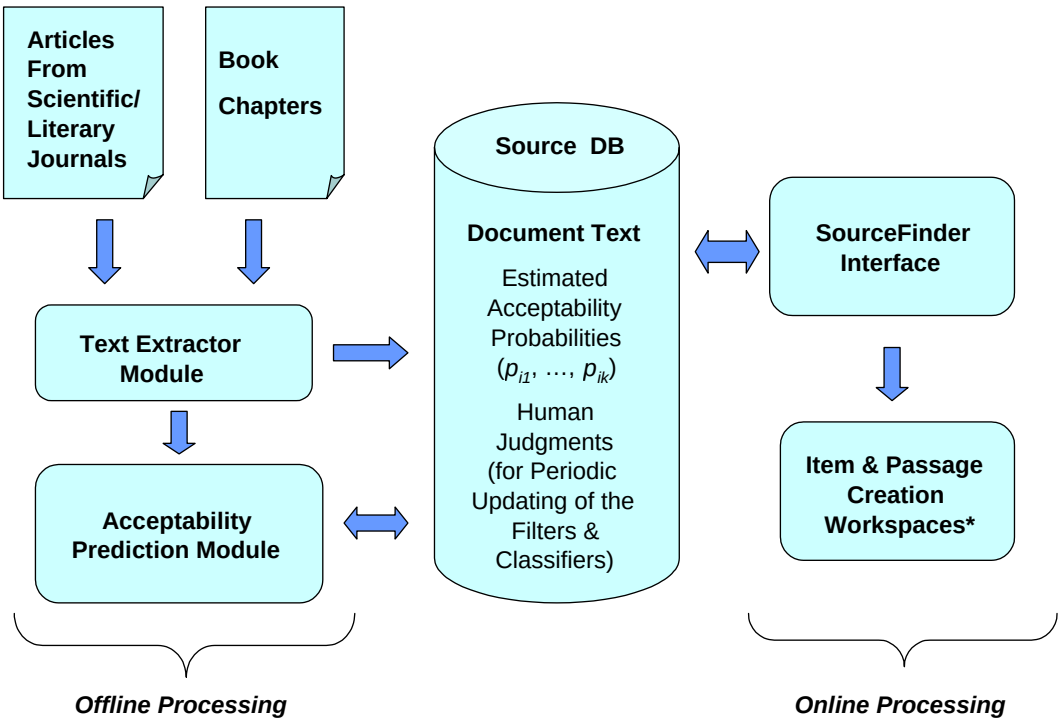
The June 2003 revision included two major design changes. First, since the strategy of downloading documents from linked pages and sites had not produced a large number of acceptable sources, the Web-crawling aspects of the system were eliminated, and the document-retrieval subsystem was redesigned to gather potential source documents from a specified set of online repositories (e.g., databases of literary and scientific journal articles.) Second, an updated document-screening capability was implemented. This updated capability was designed to provide explicit recognition of the fact that, at ETS, passage development is an on-demand activity. That is, continuous monitoring of pool status yields detailed information about the specific *types* of passages that are currently in demand, and this information is passed to item writers in the form of detailed passage-creation *assignments*. Since a given source document could be simultaneously rated as *acceptable* for one passage-creation assignment and *unacceptable* for a second passage-creation assignment, and since the number of possible passage-creation assignments is finite, the new document-screening capability was designed to provide k estimates of source acceptability for each document: one for each of k predefined source-finding assignments. This paper provides a detailed description and evaluation of this new capability. The evaluation is specified in terms of six particular GRE source-finding assignments. Results for several additional assignments are described in Sheehan et al. (2006).

The remainder of this paper is organized as follows. The first section presents an overview of the SourceFinder system that was implemented in June 2003. The second section describes the statistical approach implemented to develop source-acceptability models for the selected GRE source-finding assignments. These six models are then evaluated in the third section. The three final sections present conclusions, limitations, and directions for future research.

Incorporating Source-Acceptability Screening Into the SourceFinder System

Figure 1 presents a high-level overview of the SourceFinder system design. Two different types of components are shown: offline components (i.e., components that perform the offline

processing needed to prepare candidate source documents for operational consideration) and online components (i.e., components that facilitate real-time authoring of passages and items).



*** Has not yet been implemented.**

***Figure 1*. A high-level overview of the SourceFinder system design.**

The processing implemented within each of the components shown in Figure 1 can be summarized as follows:

1. The *Text-Extractor Module*. This component downloads candidate source documents from designated online repositories, such as databases of scientific and literary journals. All of the downloaded documents are stored in the Source Database (Source DB).
2. The *Acceptability-Prediction Module*. This module assigns a vector of source-acceptability probabilities (p_{i1}, … p_{ik}) to each document, where p_{ik} represents the probability that document *i* is acceptable for use in satisfying passage-creation assignment *k*. These probabilities are also stored in the Source DB.

3. The *SourceFinder Interface*. Test developers use this component to enter information about the specific source-finding assignment at hand and to view lists of candidate source documents sorted from most acceptable to least acceptable for the specified source-finding assignment. Retrieved documents may be saved for future reference, or may be immediately forwarded to the Item- and Passage-Creation workspaces.

4. *The Item- and Passage-Creation Workspaces*. These planned workspaces are designed to facilitate the flow of text from source, to passage, to item, and to provide a common launching site for other item-development tools.

Generating Source-Acceptability Probabilities

In many test-development areas within ETS, the search for acceptable source material is implemented in a series of stages, as follows: First, one or two test developers are charged with the task of creating a collection of candidate source documents that meet minimum acceptability requirements. An appropriate collection is then created by systematically evaluating each article in a targeted set of journals and magazines. Articles deemed acceptable are added to the collection; articles deemed unacceptable are not. The resulting collection is then made available to other test developers for use in completing specific passage-creation assignments.

The SourceFinder module implemented in June 2003 was designed to automate this process by providing a statistically based approach for distinguishing between (a) documents rated as having a relatively high probability of meeting the minimum acceptability standards specified for at least one passage-creation assignment, and (b) documents rated as being unacceptable for use in completing *any* passage-creation assignment. This goal was accomplished by providing two new capabilities: (a) a source-acceptability prediction model that assigns a vector of acceptability probabilities to each candidate source document and (b) a capability for efficiently sorting candidate documents so that test developers can easily restrict their attention to only those documents rated as having a relatively high probability of being acceptable for use in the particular source-finding assignment at hand.[1]

The steps involved in creating these capabilities for a particular testing program can be summarized as follows:

1. The program's source-finding requirements are translated into a finite set of passage-creation assignments.
2. Text characteristics that are useful for distinguishing appropriate and inappropriate documents relative to those assignments are determined.
3. Natural language processing (NLP) techniques are used to score the identified characteristics.
4. Classification and regression models are developed to predict the acceptability status of individual documents relative to each of the specified assignments. If $y_{ik} = 1$ represents the event that source i has been deemed acceptable for use in satisfying passage-creation assignment k, then the quantity being estimated is p_{ik}, the probability that $y_{ik} = 1$ (accept), rather than 0 (reject).
5. The estimated acceptability models are applied to each of the candidate source documents in the Source DB, and a vector of assignment-specific acceptability probabilities is generated for each document. These probabilities are also stored in the Source DB.
6. The graphical user interface (GUI) is amended so that test developers can request lists of candidate source documents sorted from most acceptable to least acceptable for any of the specified passage-creation assignments.

This processing is depicted graphically in Figure 2. Additional information about the processing implemented at successive stages of the analysis is summarized later in this report.

Stage 1: Feature Extraction

In this initial stage, NLP techniques are used to extract a vector of text features from each candidate source document. Each feature is designed to characterize document standing with respect to one or more of the following aspects of text variation: level of argumentation, sensitivity, accessibility, content, and subcontent. These five aspects of text variation were observed to be important determinants of source acceptability in previous research reported in Passonneau et al. (2002).

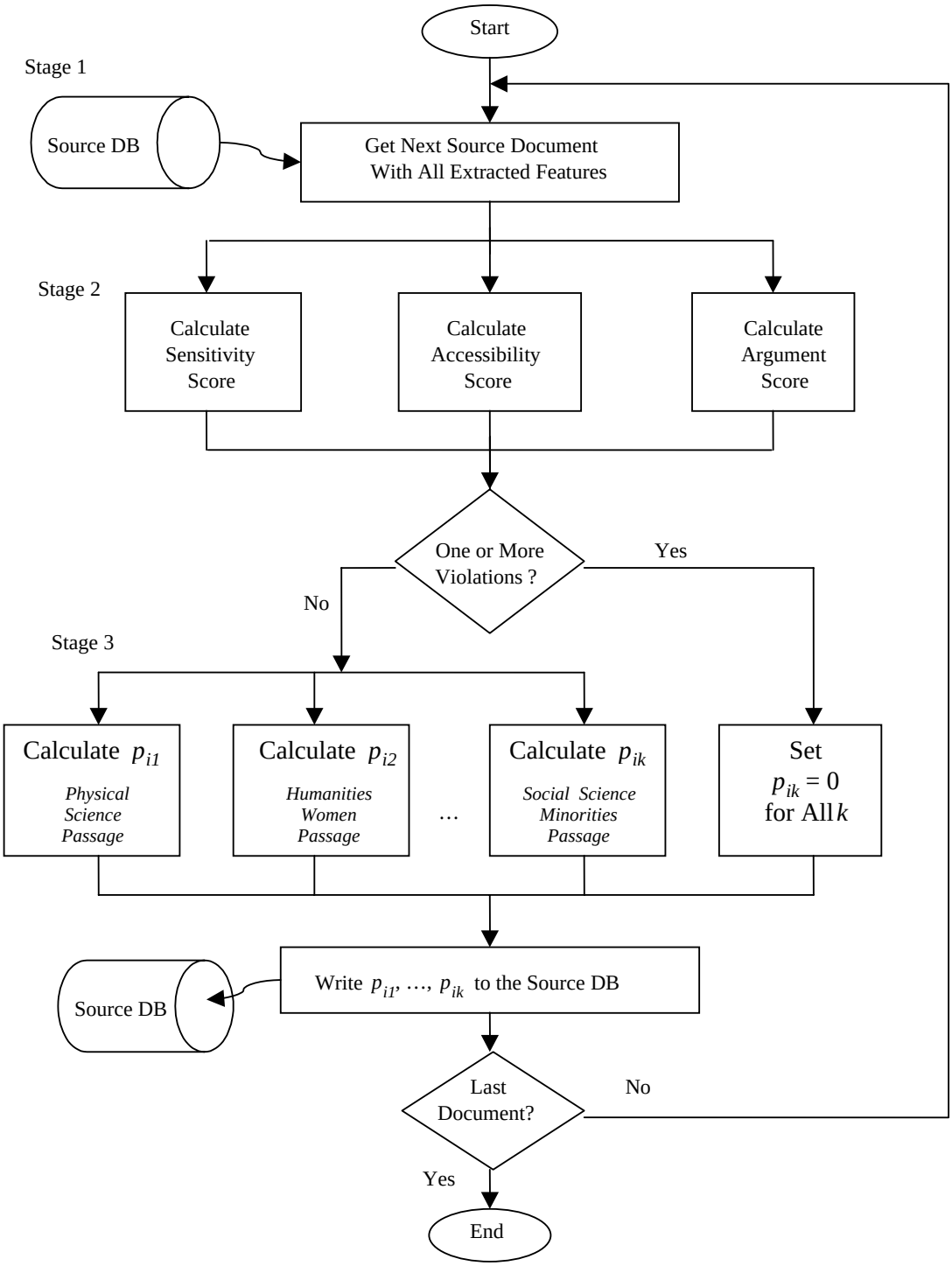


Figure 2. **Incorporating source evaluation into SourceFinder's offline processing.**

Note. Feature extraction is implemented at Stage 1, gross violations of the acceptability standards are detected at Stage 2, and more subtle variations in source acceptability are predicted at Stage 3. Final estimates are then loaded into the Source Database.

The Level of Argumentation features are designed to ensure that accepted documents offer a level of intellectual complexity that is sufficient to support one or more complex reading items. Texts that are primarily descriptive, or that merely present straightforward exposition or narration are less likely to satisfy this requirement, while texts that provide some conflict or contrast of ideas, and some uncertainty about conclusions or outcomes, are more likely to satisfy this requirement. The Sensitivity features are designed to ensure that documents rated as acceptable for a particular assignment are not in violation of ETS sensitivity guidelines. The Accessibility features are designed to ensure that accepted documents do not include specialized technical jargon that might unfairly advantage examinees with certain highly specialized backgrounds. And finally, the Content and Subcontent features are designed to ensure that source documents rated as acceptable for a particular content area actually do focus on topics appropriate to that content area. Since many of the extracted features interact with content area, the content features are also used to set content-specific threshold values.

Stage 2: Document Filtering

In this stage, a filtering process is used to detect documents that exhibit gross violations of any of the following acceptability standards: the sensitivity standard, the accessibility standard and the level of argumentation standard. As is indicated in Figure 2, documents rated as unacceptable at this stage are assigned an acceptance probability of 0 for all possible passage-creation assignments and are excluded from further processing. Documents that make it past this initial filtering stage are then processed via a logistic regression model, as described below.

Stage 3: Acceptability Prediction via Logistic Regression

One limitation of the filtering process implemented at Stage 2 is that the various acceptability standards are considered independently. This step provides a more refined prediction of acceptability status by using a logistic regression approach to evaluate the simultaneous effects of all relevant dimensions (i.e., sensitivity, accessibility, level of argumentation, and content). In particular, the acceptability status of the i^{th} document relative to the k^{th} assignment is estimated as follows:

$$p_{ik} = P(y_{ik} = 1 \mid X_{i1}, \ldots, X_{ir}) = \frac{\exp\left(\sum_{j=1}^{r} \beta_{jk} X_{ij}\right)}{1 + \exp\left(\sum_{j=1}^{r} \beta_{jk} X_{ij}\right)} + \varepsilon_{ik} \quad (1)$$

where X_{ij}, for j = 1, .., r, represent text features that are automatically extracted from each candidate source document, and the β_{jk} are coefficients that are estimated from the available training data.

Planned Redundancy

At the completion of the three processing stages described above, each document is characterized in terms of a set of *k* probability values, one for each of *k* predefined passage-creation assignments. Note that this processing involves an element of planned redundancy. In particular, both the filtering process implemented at Stage 2 and the logistic regression modeling implemented at Stage 3 are designed to draw on the *same* pool of text features and to generate predictions about the *same* aspect of variation (i.e., the probability that a given source would be judged by a trained human rater as acceptable for use in satisfying a particular passage-creation assignment). This redundant modeling strategy is designed to circumvent a problem that is inherent in the step-wise feature-selection methodology that underlies the logistic regression modeling approach. In this methodology, the available predictors compete with one another (in statistical terms) to enter the final model. The winners are determined largely by the *frequency* of the targeted violation, rather than by the *seriousness* of the targeted violation. For example, consider an extremely serious violation that only appeared once in the training sample. A text feature designed to detect this violation would have a relatively low probability of being selected for inclusion in the final logistic regression model simply because of the violation's lack of representation in the training sample. This problem is also likely to be exacerbated by multicollinearity with other features. The redundant modeling strategy described above is designed to circumvent this problem by including individual filters focused on detecting individual violations that are serious yet not frequently observed. As a consequence, unacceptable documents have two chances of being detected: The first chance is afforded by the filtering process, and the second chance is afforded by the logistic regression process.

Figure 3 illustrates how this new source-screening capability was incorporated into SourceFinder's online processing. As is indicated in the diagram, information about the specific passage-creation assignment at hand is entered via the GUI. This information includes the type of assignment (e.g. a physical science passage), the desired source length (e.g., 1,500 words), and optionally, a list of key words. Any word or phrase could be included in the list of key words. Output is then returned as a list of candidate source documents sorted from most acceptable to least acceptable for the specified assignment. The source-acceptability probabilities generated by the *Acceptability Prediction Module* constitute the data values considered by the sorting algorithm. Thus, the estimated source-acceptability probabilities help test developers focus their attention on only those documents that have a good chance of being acceptable for use in the particular source-finding assignment at hand.

Model Development for Six Specific Acceptability Models

Data

At the time that this research was conducted, the Source Database (*Source DB* in Figure 1) included more than 30,000 documents extracted from a set of 30 scientific and literary journals that had previously been used to develop GRE passages and items.[2] A subset of 136 documents extracted from this database constituted one portion of the text collection used for model development and evaluation. Each document in this subset was independently rated by two different test developers. All of the test developers participating in this portion of the study had had recent experience developing and/or reviewing passages for GRE.

The test developers provided two different types of outputs: detailed descriptions of the aspects of source variation that contributed to their acceptability decisions and numeric ratings of source acceptability expressed on a 1 – 5 scale, where 1 = Definitely Reject, 2 = Probably Reject, 3 = Uncertain, 4 = Probably Accept, and 5 = Definitely Accept. This five-point scale was used instead of a two-point (i.e., Accept/Reject) scale because the test developers thought that it would help to speed up the rating process. In particular, hard-to-rate documents could be assigned a 2, 3, or 4 rating, rather than a 1 or 5 rating. The resulting ratings were then collapsed to form a two-point Accept/Reject scale as follows: If at least one of the two test developers rated a document as either a 4 or a 5, the document was classified as an Accept; otherwise, the document was classified as a Reject.

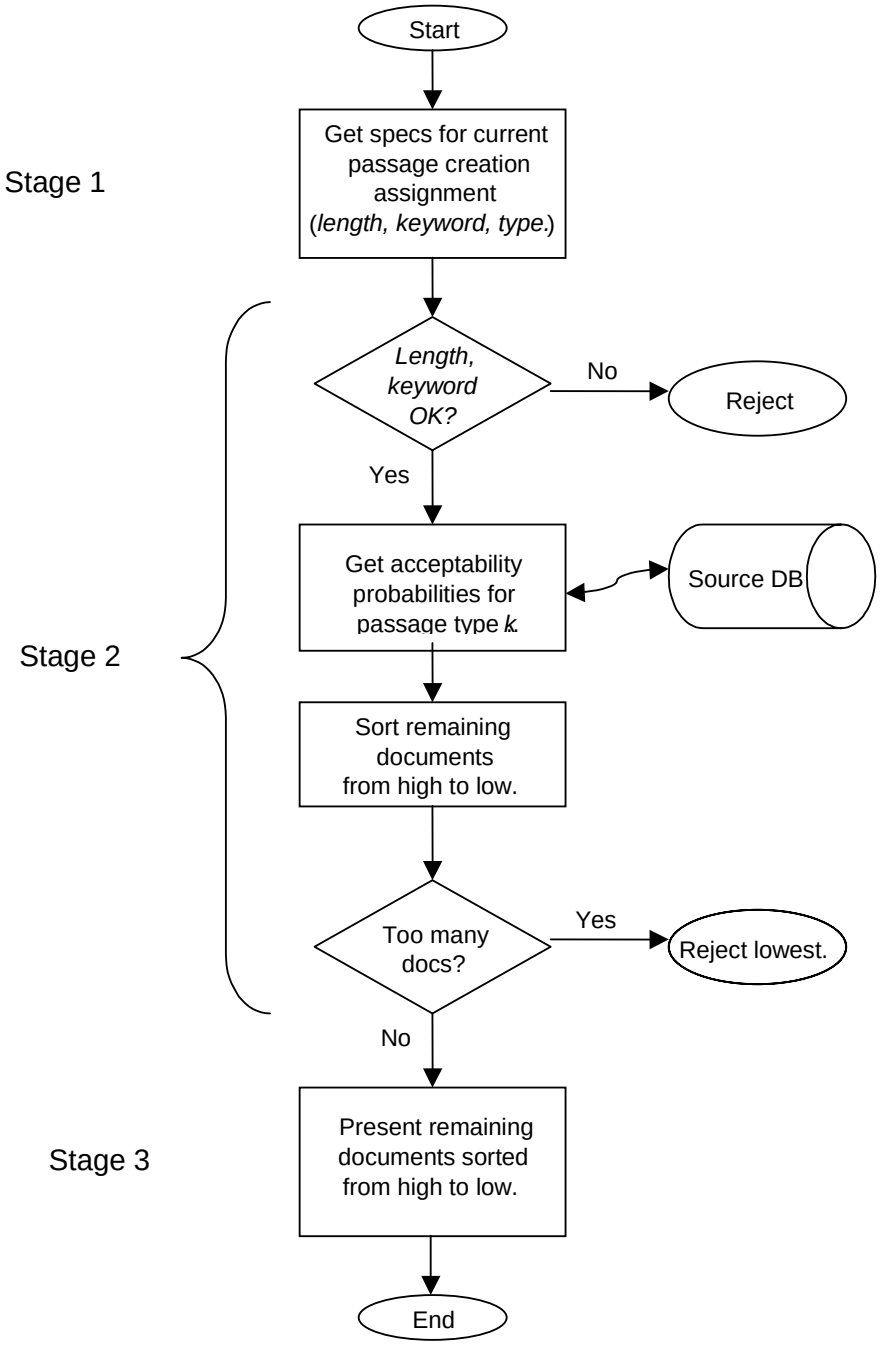


Figure 3. **Incorporating new source-screening into SourceFinder's online processing.**

Note. The specific passage-creation assignment at hand is input at Stage 1. Documents appropriate for use in completing that assignment are selected at Stage 2. Output is presented at Stage 3.

Note that the resulting Accept/Reject classifications are analogous to those that are traditionally developed in the early stages of passage evaluation. In particular, an Accept classification is an indication that a document should be retained for consideration by other test developers, because it's likely that at least one test developer will be able to use the document to complete at least one passage-creation assignment. Note that this does not mean that the document will be acceptable to *all* test developers.

The classification results obtained for the 136 documents rated by the GRE test developers are shown in Table 1. As can be seen, 92 of the 136 sources were rejected, while 44 were accepted, yielding an acceptance rate of about 32%. This rate is similar to the acceptance rate yielded by the interactive document-screening capability included in the previous SourceFinder module (see Passonneau, et al., 2002). Also, since only 44 acceptable sources were found, the database was augmented with a set of 86 historical sources; that is, source documents that had previously been used as sources for operational GRE reading-comprehension sets. This raised the number of acceptable sources to 130, or slightly more than half of the available documents.

Table 1

Source Documents for Use in Training and Evaluation

Document type	Accept	Reject	Total
Potential new sources: Documents downloaded from designated online scientific and literary journals	44	92	136
Historical sources: Documents previously used to create operational GRE passages and items	86	0	86
Total	130	92	222

Passage-Creation Assignments

The source-acceptability models developed for six particular GRE assignments are documented in this report. These six assignments are shown in Table 2. Each assignment is specified in terms of a particular content area, a particular passage length, and in two cases, a particular subcontent area.

Table 2

Passage-Creation Assignments

No.	Content area	Passage length (in words)	Subcontent Area	Label
1	Physical Science	150	Any	PS
2	Biological Science	150	Any	BS
3	Social Sciences	150	Any	SS
4	Humanities	150	Any	HU
5	Social Science	150	Women	SS-Wo
6	Humanities	150	Women	HU-Wo

Three Stages of Model Development

As was noted previously, model development was implemented in stages. More than 50 different features were considered at Stage 1 (feature extraction). Each feature was designed to capture variation in one or more of the five dimensions of source acceptability described above (e.g., Level of Argumentation, Sensitivity, Accessibility, Content, and Subcontent). Additional information about the individual features considered at this stage of the analysis is provided in the section headed *Sample Features* in this report.

Second, filters for use in detecting gross violations of the Sensitivity, Accessibility, and Level of Argumentation standards were developed. A total of 19 different features were selected for inclusion in one or more of the three filters developed at this stage. Third, a logistic regression approach, implemented via maximum likelihood estimation with iteratively reweighted least squares (Hastie, Tibshirani, & Friedman, 2001), was used to develop a distinct source-acceptability model for each assignment. A total of 20 different features were selected for inclusion in one or more of the estimated equations. This set included some features that had also been selected for inclusion in the Level of Argumentation, Sensitivity, or Accessibility filters.

Sample Features

Table 3 lists sample features for each of the five dimensions of source variation considered in this study. Each feature in the table was found to be significantly related to variation in test developers' ratings of source acceptability ($p < .01$) for one or more of the

assignment-specific acceptability models documented in this report. A brief description of each feature is provided below.

Table 3

Sample Features by Dimension of Source Variation

Dimension of variation addressed	Sample features
Level of argumentation	Number of academic verbs per 1,000 words (e.g., suggest, consider, indicate) Number of academic conjuncts per 1,000 words (e.g., alternatively, however)
Sensitivity	Number of red flag words per 1,000 words (e.g., abortion, amputated) Number of yellow flag words per 1,000 words (e.g., addicted, depressed)
Accessibility	Number of nominalizations per 1,000 words (e.g., assumption, amazement) Type-token ratio
Content	Similarity to the physical science target vector Similarity to the biological science target vector
Subcontent	Number of female pronouns per 1,000 words (e.g., she, her, hers, herself) Similarity to the female target vector

Level of Argumentation Features

Source documents for GRE reading-comprehension passages must exhibit a level of argumentation that is sufficient to support one or more difficult reading questions. There are many different ways that a text might fail to meet this requirement. For example, a text that only

presents one side of an argument is less likely to be acceptable, while texts that consider multiple viewpoints are more likely to be acceptable. Similarly, texts that are primarily descriptive, or that merely present straightforward exposition or narration, are less likely to be acceptable, while texts that present a degree of conflict or uncertainty are more likely to be acceptable.

Table 3 lists two features that were found to be of use in distinguishing between texts with acceptable and unacceptable types of argumentation: the Number of Academic Verbs per 1,000 words, and the Number of Academic Conjuncts per 1,000 words. These two features are based on previous research published in Biber (1988) and in Biber, Johansson, Leech, Conrad, and Finegan (1999). The list of academic verbs includes verbs like *suggest*, *consider*, and *indicate* that, according to Biber et al. (1999), tend to occur more frequently in academic texts and less frequently in nonacademic texts. The list of academic conjuncts includes conjuncts like *alternatively*, *consequently*, and *however* that, according to Biber (1988), tend to occur more frequently in academic texts and less frequently in nonacademic texts. Figure 4 provides an indication of how well these features performed relative to the task of distinguishing a particular type of level of argumentation violation, i.e., text that is too narrative to support the type of academic argumentation that is typically presented in GRE passages. The top plot presents results for the Academic Verbs feature; the bottom plot presents results for the Academic Conjuncts feature. Each plot illustrates the range of feature variation observed for 138 documents. This set includes three different types of documents:

1. Documents that had been rated as unacceptable for use in GRE passage development because they were too narrative (n = 8)
2. Documents that had been rated as acceptable for use in GRE passage development (n = 44)
3. The set of all available historical sources (n = 86)

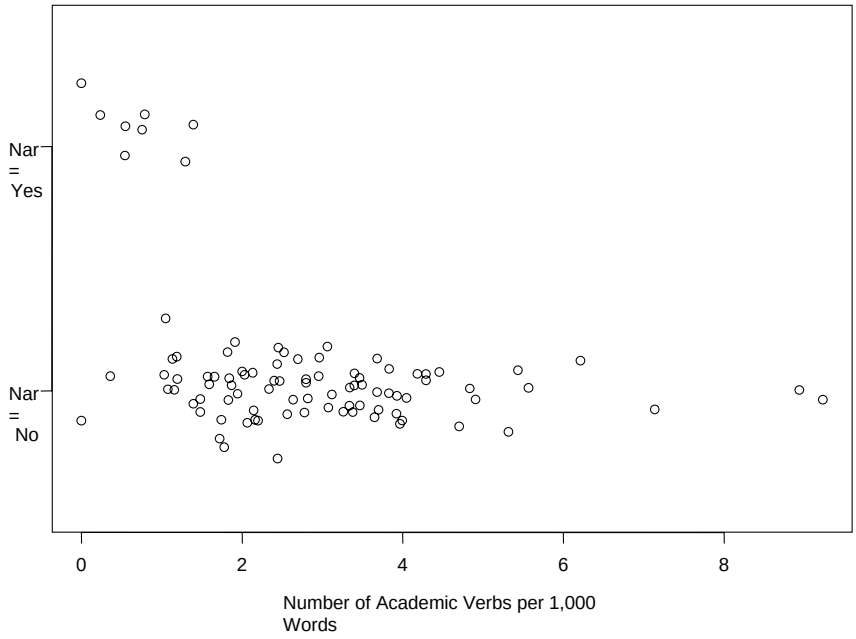
The vertical axis in each plot indicates the true Level of Argumentation classification recorded for each document. The eight documents classified as being too narrative are plotted at the point labeled *Nar = Yes*. The 130 documents classified as providing an acceptable Level of Argumentation (including both the 44 documents rated as Acceptable and the 86 historical sources) are plotted at the point labeled *Nar = No*, indicating that no narrative violations were

detected. The small amount of vertical scatter shown in each set of points was created by adding a small amount of random noise to each point's *y*-value so that points that would otherwise be plotted on top of each other appear at distinct vertical locations. This technique is called *vertical jittering* (Chambers, Cleveland, Kleiner, & Tukey, 1983). Vertical jittering is frequently used to enhance the interpretability of two-dimensional scatter plots when the variable plotted on the vertical axis is measured on a discrete scale (e.g., accept/reject, yes/no, etc.). The plots in Figure 4 suggest that, for each feature, a low value is an indication that the targeted violation might be present (i.e., the text might be too narrative) and a high value is an indication that that the targeted violation most likely is *not* present.

Sensitivity Features

Variation with respect to the sensitivity dimension of source acceptability is measured via two different types of sensitivity word lists. The first list is called the *Red Flag List*. It includes words and phrases that have a high probability of being present in documents rated as containing sensitivity violations and a low probability of being present in documents rated as *not* containing sensitivity violations. The second list is called the *Yellow Flag List*. It contains words and phrases that are only moderately useful for detecting texts containing sensitivity violations. The sensitivity features shown in Table 3 capture variation in the sensitivity dimension of source acceptability by counting the number of red flag and yellow flag words detected in each candidate source document.

Accessibility features. The accessibility dimension of source acceptability refers to the probability that a particular text might unfairly advantage an examinee with some specialized background knowledge, for example, detailed knowledge of cell biology. Table 3 lists two features that were found to be useful in measuring this particular aspect of source acceptability. The first feature is the Number of Nominalizations per 1,000 words. This is a count of the number of words in the document that ended in any of the following suffixes: *-tion*, *-ment*, or *-ity*. The second feature is based on previous research reported in Youmans (1991). This research demonstrated that certain vocabulary-usage characteristics may be measured by comparing the total number of unique words in a document to the absolute number of words in the document.



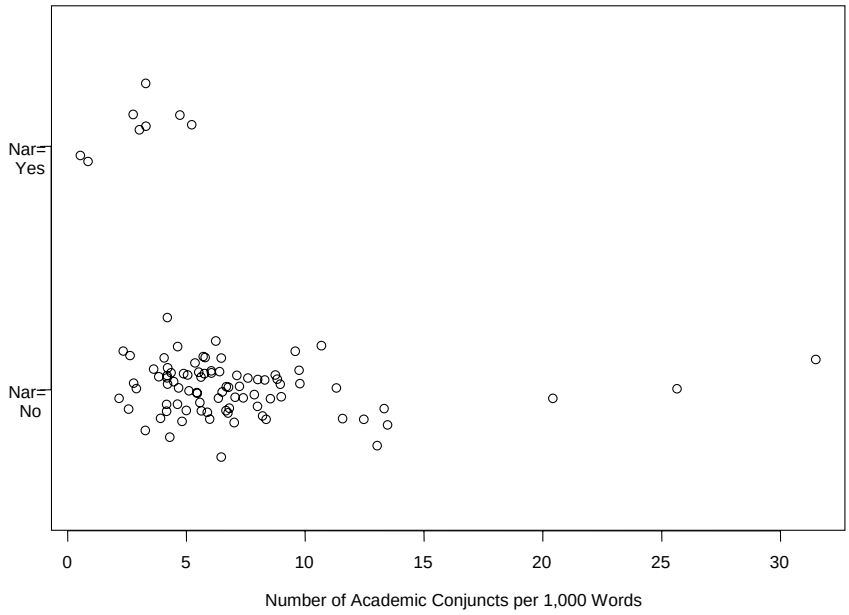


Figure 4. **Two of several features designed to characterize document standing relative to the Level of Argumentation dimension of source variation.**

The first quantity (i.e., the total number of unique words in a document) is often referred to as the number of word *types*. The second quantity (i.e., the absolute number of words in a document) is often referred to as the number of word *tokens*. Youmans demonstrated that the ratio of word types to word tokens, also called the *type-token ratio*, is useful for distinguishing texts that employ relatively diverse vocabularies. The current analyses demonstrated that documents rated as too jargony tended to have relatively high values on both the nominalizations feature and the type-token ratio. Consequently, these features are used to distinguish texts that contain a large amount of jargon and, as a result, are likely to be rated as exhibiting an accessibility violation.

Content features. Because the specific content area to be addressed by a given passage is specified in the item writer's assignment, and because a large number of text features could potentially interact with content area, accurate assessment of source content is an important component of source acceptability. SourceFinder evaluates the content dimension of source acceptability by implementing a *content vector approach* (Salton, 1989; Salton & McGill, 1983).

Content vector approaches have been successfully applied in a variety of assessment contexts, including, for example, *e-rater*®, the automated essay-scoring system developed at ETS (Burstein, 2003; Burstein et al., 1998). Table 4 highlights some similarities and differences between the content vector approach implemented in the current application and the approach implemented in the *e-rater* application.

As is indicated in Table 4, the content vector approach implemented in the *e-rater* application is designed to assess degree of similarity between the content words present in training essays scored by trained human raters and the content words present in student essays submitted for scoring. In a similar vein, the content vector approach implemented in the current application is designed to assess degree of similarity between the content words present in training passages classified by expert test developers and the content words present in new, unrated documents extracted from the Source Database. Further, in each approach, both the classification categories of interest, and the individual documents to be classified are represented as vectors in a multidimensional vector space. The primary advantage of this particular representational strategy is that it admits a useful, easily implemented similarity measure; i.e., the cosine similarity measure (Salton, 1989; Salton & McGill, 1983).

Table 4

Use of Content Vector Analyses in e-rater and in SourceFinder

Text to be classified	Classification categories	Similarity measures
e-rater		
A new, unrated essay	1. Training essays with a human rater score of 1 2. Training essays with a human rater score of 2 3. Training essays with a human rater score of 3 4. Training essays with a human rater score of 4 5. Training essays with a human rater score of 5 6. Training essays with a human rater score of 6	Cosine between the vector of normalized term frequencies estimated for the new, unrated essay, and the vectors of normalized term frequencies estimated for each classification category
SourceFinder		
A new, unrated journal article or book chapter	1. Previously developed passages classified as appropriate for use in satisfying the humanities content specification 2. Previously developed passages classified as appropriate for use in satisfying the social sciences content specification 3. Previously developed passages classified as appropriate for use in satisfying the physical sciences content specification 4. Previously developed passages classified as appropriate for use in satisfying the biological sciences content specification	Cosine between the vector of normalized term frequencies estimated for the new, unrated potential source document, and the vectors of normalized term frequencies estimated for each classification category

The content vector approach developed for the current application included six steps, as summarized below.

1. A target text was constructed to represent each targeted content category (e.g., the Physical Sciences content category, the Biological Sciences content category, etc.). The target text for content area *k* was obtained by concatenating together previously developed passages from content area *k*. A total of 261 previously developed GRE passages were considered at this stage of the analysis.

2. Both the *k* target texts and each of the candidate source texts in the Source Database were represented as vectors of normalized word frequencies. Each vector contained *w* frequency values, one for each of the *w* content words selected for consideration in the analyses. Selected content words included all of the nouns, verbs, adjectives, and adverbs detected in at least two of the 261 passages.

3. Because the resulting vectors were quite long, two different approaches for collapsing across rows indexed by similar content words were implemented. First, a stemming tool was used to collapse across rows associated with words arising from a common word stem. For example, *reading*, *read*, and *reads* were each represented by the single word class *to read*. Second, a measure of word-word similarity (Lin, 1998) was used to collapse across words rated as having a high degree of distributional similarity. The Lin similarity measure uses word co-occurrence frequencies extracted from a large corpus of natural language text to assess word-word similarity. The approach is based on Harris' (1968) distributional hypothesis, which states that words with similar meanings tend to appear in similar contexts. For example, note that the words *bacteria* and *germs* are frequently used with the following verbs

   grows, lives, spreads, and causes

   and are often modified by the following adjectives

   harmful, air borne, and *deadly.*

   Based on this type of corpus evidence, *bacteria* and *germs* were rated as having a high degree of distributional similarity and were subsequently collapsed into a single word class. Note that the resulting collapsing strategy preserves part-of-speech

information. That is, nouns are only collapsed with other nouns, verbs are only collapsed with other verbs, and adjectives and adverbs are only collapsed with other adjectives and adverbs. It is also useful to note that the approach includes, but is not limited to, collapsing across close synonyms.

4. In this step, the word classes defined above are viewed as distinct dimensions of variation and the resulting frequency vectors are viewed as observations in *t*-space, where $t < w$ is the total number of word classes remaining at the completion of the collapsing algorithm. The degree of similarity between the i^{th} source document and the k^{th} target text is then estimated as follows:

$$r_{ik} = \frac{\sum_{j=1}^{t} s_{ij} g_{kj}}{\left(\sum_{j=1}^{t} s_{ij}^2 \sum_{j=1}^{t} g_{kj}^2 \right)^{\frac{1}{2}}} \qquad (2)$$

where $S_i = [s_{i1}, \ldots, s_{it}]$ is the collapsed vector of normalized term frequencies obtained for the i^{th} source document, and $G_k = [g_{k1}, \ldots, g_{kt}]$ is the collapsed vector of normalized term frequencies obtained for the k^{th} target text. This measure is called the *cosine similarity measure* or the *cosine correlation* because it is mathematically equal to the cosine of the angle between $S_i = [s_{i1}, \ldots, s_{it}]$ and $G_k = [g_{k1}, \ldots, g_{kt}]$. That is, if θ represents the angle between S_i and G_k, then $r_{ik} = \cos(\theta)$.

This particular similarity measure is frequently used in text-classification applications, including, for example, the *e-rater* application, because it is known to be relatively insensitive to zero-frequency word classes (Leydesdorff, 2005). That is, the absence of a particular word class (e.g., *bacteria*|*germ*) does not indicate dissimilarity as strongly as the presence of a matching word class indicates similarity. This property is particularly desirable for the current application because many of the word classes included in the *k* target vectors have only a small probability of being present in any new document.

Selected portions of the target vectors developed for the four main GRE content areas are shown in Table 5. Individual word classes found to be indicative of particular content categories are shaded. The results suggest that the GRE content areas tend to employ relatively distinct vocabularies. For example, words like *species*, *population*, *brain*, *process*, *bacteria*, and *germ* tend to occur with relatively high frequency in biological science texts, and relatively low frequency in the other three types of texts. Similarly, words like *art*, *work*, *literary*, *artistic*, *writer*, and *novel* tend to occur with relatively high frequency in humanities texts, but relatively low frequency in each of the other three types of texts.

It is important to note, however, that the content categories listed in Table 5 are not mutually exclusive. For example, it is possible for a given source document to be appropriate for use in developing *either* a social science passage *or* a humanities passage. Similarly, some science documents are consistent with *both* the physical science content category and the biological science content category.

The utility of this approach for predicting the content dimension of source acceptability is illustrated in Figure 5. The figure shows the cosine similarity measures obtained for a set of 68 training documents that had been classified by GRE test developers as having content that would be appropriate for use in developing a GRE physical science passage. This dataset provides an independent validation of the approach because it does not include any of the 261 passages used to develop the *k* target vectors used for model development. The figure shows that, for all but three of the documents, the similarity to the physical science content vector exceeded that for each of the other three content area vectors (i.e., biological science, social science, and humanities.) An examination of the three unusual documents revealed that, in all three cases, (a) the test developers had indicated that the document was appropriate for use as *either* a physical science passage *or* a *b*iological science passage, and (b) the cosine similarity estimates obtained for the physical science content vector and the biological science content vector were quite close. Thus, although all three of the documents were observed to be most highly correlated with the biological science content vector, each also yielded a high correlation with the physical science content vector. This confirms that the approach yields useful information about the content areas addressed by candidate source documents extracted from the Source Database. In addition, the fact that independent estimation and validation datasets were used suggests that the results are likely to generalize to any new sample of documents drawn from the Source Database.

Table 5

Normalized Term Frequencies for Selected Word Classes (Frequency per 1,000 Words)

Word class	Biological science	Physical science	Social science	Humanities
Species (N)	3.69	0.33	0.15	0.03
Population (N)	2.78	0.00	0.93	0.00
Brain (N)	2.01	0.00	0.00	0.00
Process (V)	1.75	0.92	0.06	0.21
Bacteria|Germ (N)	1.49	0.00	0.00	0.00
Surface (N)	0.19	4.29	0.03	0.18
Earth (N)	0.39	4.09	0.00	0.07
Star (N)	0.00	3.83	0.00	0.00
Planet (N)	0.00	3.10	0.00	0.00
Electron|Neutron|Particle (N)	0.00	1.91	0.00	0.07
Political|Ideological (A)	0.06	0.13	3.06	0.57
Societal|Social (A)	0.00	0.00	3.00	0.75
Historian (N)	0.00	0.00	2.85	0.50
Class (N)	0.00	0.00	1.86	0.64
Movement (N)	0.06	0.13	1.74	0.57
Art (N)	0.00	0.00	0.03	4.41
Work (N)	0.00	0.20	2.13	3.06
Literary|Artistic (A)	0.00	0.00	0.09	2.88
Writer (N)	0.00	0.00	0.09	2.74
Novel (N)	0.06	0.00	0.00	2.49

Note. Letters in parentheses indicate part of speech, as follows: N = noun, V = verb, A = adjective or adverb. The construction *Word1*|*Word2* indicates words classified as having a high degree of distributional similarity.

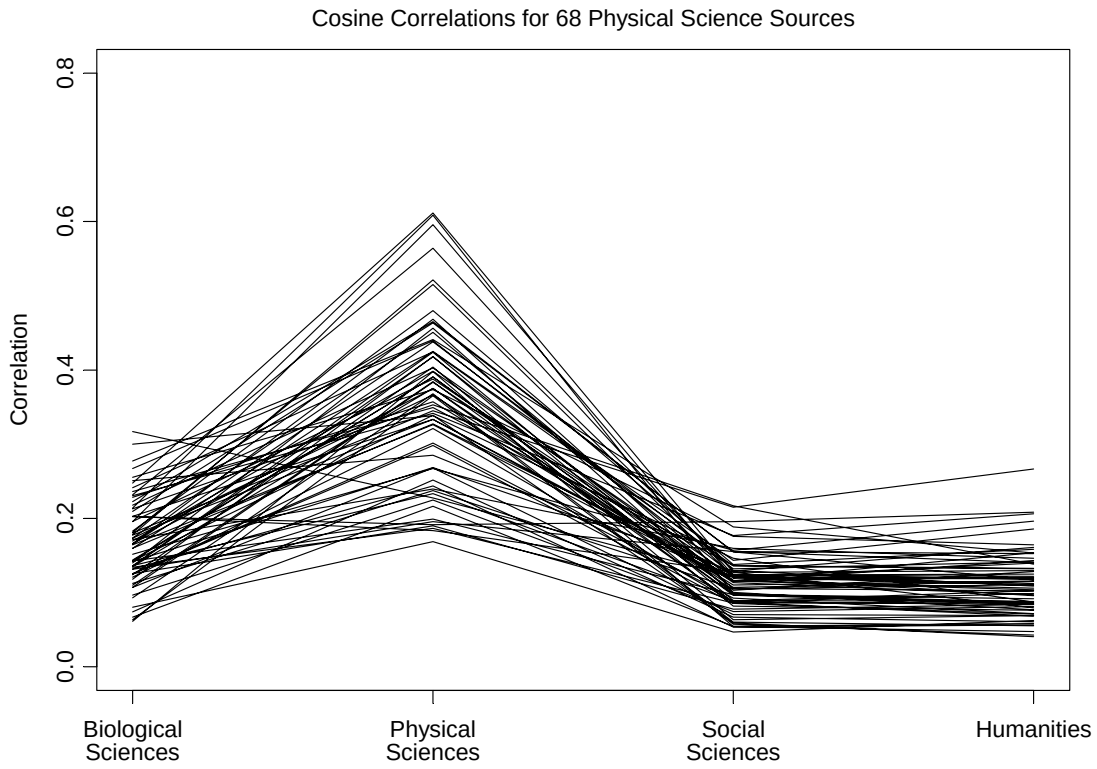


Figure 5. **Using a content vector approach to predict the content area addressed by 68 documents classified by expert test developers as exhibiting content appropriate for use in constructing a GRE physical science passage.**

Subcontent features. Table 3 also lists two features that were found to provide useful information about the subcontent dimension of source acceptability. The first feature, similarity to the Female Target Vector was constructed by implementing a content vector approach as described above. The second feature was constructed by counting the number of "female" pronouns (e.g., she, her, hers, herself) in each candidate source text.

Model Evaluation

Two separate evaluations are presented in this section: (a) an evaluation of the filters implemented at Stage 2, and (b) an evaluation of the logistic regression models implemented at Stage 3.

An Evaluation of the Filters Implemented at Stage 2

The filtering process focuses on four specific dimensions of source variation: length, sensitivity, accessibility, and level of argumentation. As was previously noted in Figures 2 and 3, the length filter is implemented as part of SourceFinder's online processing, while each of the other three filters are implemented as part of SourceFinder's offline processing.

Implementation of the length filter is completely straightforward: The user specifies a desired source length, and only those documents that meet or exceed that length are included in the set of documents returned. This approach is designed to insure that all returned documents are long enough to support the specific passage-creation assignment at hand. Test developers can also use this capability to insure that all returned documents are long enough to provide a passage of the desired length without violating copyright restrictions.

The current SourceFinder module also includes three offline filters: a sensitivity filter, an accessibility filter, and a level of argumentation filter. These offline filters are designed to detect violations that are rare, yet serious enough to warrant immediate exclusion from further processing. Exploratory data analyses and tree-based classification techniques (Brieman, Friedman, Olshen, & Stone, 1984) were used to develop these filters.

The basic form of the sensitivity filter is shown in Figure 6. Note that both sensitivity features and content features are considered in the filter definition. This strategy is designed to accommodate the fact that different content areas have different sensitivity requirements. For example, a humanities text containing the word *sexual* will have a higher probability of being in violation of the sensitivity standard than will a biological sciences text that contains that same word.

Precision. SourceFinder's offline filtering process may be viewed as a binary classification problem: Each potential new source document is classified as either containing or not containing any of three different types of violations: sensitivity violations, accessibility violations, or level of argumentation violations. This type of classification problem is commonly evaluated by considering the proportion of times that the classifications generated by the system are confirmed by trained human experts. In the application discussed here, the proportion of interest is the proportion of times that a document that SourceFinder classified as exhibiting a specific type of violation, say, a sensitivity violation, was also classified as exhibiting that type of violation by a trained human expert. This proportion is often referred to as the system's *precision* (van Rijsbergen, 1979). When many of the documents that SourceFinder classifies as

being in violation are also judged by expert test developers as being in violation, precision will be high; otherwise it will be low. Precision estimates reflecting training sample performance for each of SourceFinder's offline filters are shown in Table 6. The table shows, for example, that the accessibility filter detected a total of 17 accessibility violations and that, of those, 12 (i.e., 71%) turned out to be true accessibility violations, as determined from test-developer ratings. This confirms that the filter is operating as planned, since the clear majority of detected violations turned out to be true violations. Note, however, that the total number of detected violations is quite small. This indicates that additional research aimed at detecting gross violations of the sensitivity, accessibility, and level of argumentation filters is needed. Also, when considering these results, it is important to recall that the filters are designed to detect acceptability violations that are serious yet *rare*. As more and more of these rare violations are uncovered, the impact of the filtering process is likely to be more substantial.

Start
Get next document
Set Sensitivity Score equal to zero
If Content Feature 1 $\geq c_1$ *and Sensitivity Feature 1* $\geq s_1$
then set Sensitivity Score equal to 1
…
If Content Feature n $\geq c_n$ *and Sensitivity Feature n* $\geq s_n$
then set Sensitivity Score equal to 1
If Sensitivity Score = 1 then Reject; else Accept
If not last document, return to Start
End

Figure 6. **The basic format of SourceFinder's sensitivity filter. The threshold values,** c_1 **through** c_n **and** s_1 **through** s_n**, were developed through exploratory data analyses and tree-based classification techniques.**

Table 6

The Precision of SourceFinder's Offline Filters

Type of filter	Total violations detected	Number of true violations	Number of false violations	Precision
Sensitivity	6	4	2	67%
Accessibility	17	12	5	71%
Level of argumentation	12	12	0	100%

An Evaluation of the Logistic Regression Models Implemented at Stage 3

A logistic regression model was generated for each of the assignments listed in Table 2. Three separate evaluations of these models are reported. The first evaluation focuses on the document *ranks* induced by the estimated models. The second evaluation focuses on the document *classifications* induced by an acceptance rule defined in terms of the generated acceptability probabilities. The third section presents cross-validation results for a particular passage-creation assignment.

Evaluation via document ranks. As was previously indicated in Figure 3, candidate source documents are presented to SourceFinder users as a sorted list. Clearly, the greatest reductions in source-evaluation times will be achieved when all of the truly acceptable documents are sorted to the beginning of the list and all of the truly unacceptable documents are sorted to the end of the list. This suggests that the practical utility of the estimated logistic regression models may be evaluated by considering the extent to which the document ranks induced by the estimated acceptability probabilities are useful for distinguishing between documents that the test developers rated as acceptable for the given assignment, and documents that the test developers rated as unacceptable for the given assignment. The evaluation approach documented in this section is designed to provide such information.

Table 7 shows the document ranks induced by the physical science logistic regression model. The column labeled *SF* provides the acceptability probabilities generated by the physical science logistic regression model. The column labeled *TD* provides the corresponding acceptability judgments collected from the test developers (1 = acceptable for use as a physical

science source, 0 = unacceptable for use as a physical science source). Documents are sorted in terms of the generated acceptability probabilities. Thus, the document with the highest acceptability probability for the physical science assignment is listed first (e.g., Rank = 1, SF = 0.9911), and the document with the lowest acceptability probability for the physical science assignment is listed last (e.g., Rank = 222, SF = 0.0035). The table shows that the documents that the human judges rated as acceptable (i.e., TD = 1) tend to have high probabilities and low ranks, while the documents that the human judges rated as unacceptable (TD = 0) tend to have low probabilities and high ranks. This suggests that the physical science acceptability model has succeeded in capturing important test-developer requirements, and that a strategy of using these probabilities to sort documents from high to low, before presenting them to SourceFinder users, may help test developers find more high-quality physical science sources in less time.

Similar tables were developed for each of the other assignments considered in this evaluation. These tables are included in the appendix. The tables confirm that all six of the assignment-specific logistic regression models were successful at generating sets of acceptability probabilities that were useful for distinguishing between the documents that the test developers had rated as acceptable for a given assignment, and those that they had rated as unacceptable for a given assignment.

Operating characteristic curves. SourceFinder's ability to generate document ranks that are useful for distinguishing between acceptable and unacceptable documents was also evaluated by constructing an operating characteristic (OC) curve for each assignment, as follows:

1. First, the acceptability probabilities obtained for each assignment were used to order the 222 training documents from high to low; that is, from those rated as most acceptable for the given assignment to those rated as least acceptable for the given assignment.

2. The sorted documents were then divided into groups such that the first group contained the 20 documents with the highest acceptability probabilities, the second group contained the 20 documents with the next highest acceptability probabilities, an so on in that manner. Since there were 222 documents, this process yielded 12 groups: 11 groups of 20 documents and a 12th group containing just 2 documents. The 12th group was then merged with the 11th group, yielding a final set of 11 groups: 10 composed of 20 documents and 1 composed of 22 documents.

Table 7

Document Ranks Induced by the GRE Physical Science Acceptability Model

Rank	SF	TD	Rank	SF	TD	Rank	SF	TD	Rank	SF	TD
1	.9911	1	56	.4006	0	111	.0136	0	166	.0084	0
2	.9907	1	57	.3930	0	112	.0136	0	167	.0083	0
3	.9889	1	58	.3808	1	113	.0136	0	168	.0083	0
4	.9830	1	59	.3530	1	114	.0135	0	169	.0081	0
5	.9699	1	60	.3271	0	115	.0135	0	170	.0080	0
6	.9672	1	61	.2766	0	116	.0134	0	171	.0076	0
7	.9481	1	62	.2676	0	117	.0131	0	172	.0076	0
8	.9395	1	63	.2523	1	118	.0130	0	173	.0075	0
9	.9371	1	64	.2435	0	119	.0128	0	174	.0073	0
10	.9354	1	65	.2009	1	120	.0127	0	175	.0073	0
11	.9293	1	66	.1113	0	121	.0125	0	176	.0073	0
12	.9242	1	67	.0430	0	122	.0124	0	177	.0073	0
13	.9142	1	68	.0414	1	123	.0121	0	178	.0072	0
14	.9124	1	69	.0413	0	124	.0121	0	179	.0071	0
15	.9110	1	70	.0375	0	125	.0119	0	180	.0070	0
16	.8952	1	71	.0334	0	126	.0119	0	181	.0069	0
17	.8947	1	72	.0321	0	127	.0118	0	182	.0069	0
18	.8941	1	73	.0298	0	128	.0118	0	183	.0069	0
19	.8860	1	74	.0291	0	129	.0117	0	184	.0069	0
20	.8856	1	75	.0278	0	130	.0117	0	185	.0068	0
21	.8649	0	76	.0271	0	131	.0116	0	186	.0067	0
22	.8633	1	77	.0256	0	132	.0112	0	187	.0067	0
23	.8558	1	78	.0255	0	133	.0111	0	188	.0067	0
24	.8549	1	79	.0234	0	134	.0111	0	189	.0067	0
25	.8423	1	80	.0230	1	135	.0109	0	190	.0067	0
26	.8418	1	81	.0227	0	136	.0109	0	191	.0066	0
27	.8367	0	82	.0209	0	137	.0109	0	192	.0066	0
28	.8301	1	83	.0207	0	138	.0107	0	193	.0065	0
29	.8284	0	84	.0207	0	139	.0104	0	194	.0063	0

(Table continues)

Table 7 (continued)

Rank	SF	TD	Rank	SF	TD	Rank	SF	TD	Rank	SF	TD
30	.8120	1	85	.0202	0	140	.0104	0	195	.0063	0
31	.8116	1	86	.0188	0	141	.0103	0	196	.0062	0
32	.8082	1	87	.0187	0	142	.0102	0	197	.0062	0
33	.7962	1	88	.0187	0	143	.0102	0	198	.0062	0
34	.7927	1	89	.0179	0	144	.0101	0	199	.0062	0
35	.7905	1	90	.0178	0	145	.0100	0	200	.0061	0
36	.7706	0	91	.0167	0	146	.0100	0	201	.0061	0
37	.7607	0	92	.0166	0	147	.0099	0	202	.0060	0
38	.7513	1	93	.0165	0	148	.0098	0	203	.0059	0
39	.7402	0	94	.0161	0	149	.0097	0	204	.0058	0
40	.7309	1	95	.0159	0	150	.0096	0	205	.0056	0
41	.7239	1	96	.0157	0	151	.0096	0	206	.0053	0
42	.7229	0	97	.0157	0	152	.0095	0	207	.0051	0
43	.7087	1	98	.0156	0	153	.0094	0	208	.0051	0
44	.7084	1	99	.0151	0	154	.0093	0	209	.0049	0
45	.6911	1	100	.0149	0	155	.0093	0	210	.0049	0
46	.6881	1	101	.0148	0	156	.0091	0	211	.0049	0
47	.6719	0	102	.0147	0	157	.0090	0	212	.0048	0
48	.6103	0	103	.0147	0	158	.0090	0	213	.0048	0
49	.5993	1	104	.0146	0	159	.0089	0	214	.0047	0
50	.5552	0	105	.0145	0	160	.0089	0	215	.0047	0
51	.4981	1	106	.0142	0	161	.0087	0	216	.0046	0
52	.4971	1	107	.0140	0	162	.0086	0	217	.0046	0
53	.4935	0	108	.0140	0	163	.0085	0	218	.0046	0
54	.4410	0	109	.0138	0	164	.0085	0	219	.0046	0
55	.4144	1	110	.0137	0	165	.0084	0	220	.0040	0
									221	.0035	0
									222	.0035	0

3. The proportion of *TD Accepts* in each group was determined.

4. These proportions were then plotted against group rank order.

The resulting OC curves are plotted in Figure 5. Six curves are shown, one for each of the six different passage-creation assignments specified in Table 2. Individual assignments are identified with the assignment labels listed in Table 2. Thus, for example, the curve labeled *PS* is the OC curve for the physical science assignment, and the curve labeled *BS* is the OC curve for the biological science assignment. The plot confirms that, for all but one of the specified assignments, the acceptability probabilities generated by the estimated logistic regression models were successful at sorting the available documents so that the documents that the human raters had classified as acceptable appeared near the top of the list and the documents that the human raters had classified as unacceptable appeared near the bottom of the list.

The one curve that exhibited a slightly less satisfactory performance profile is the curve that was estimated for the social science (SS) assignment. The normalized term frequencies in Table 5 provide a plausible explanation for this result. The frequencies confirm that, among the various content areas considered in the analyses, only the social science content area lacks a unique vocabulary. To see this, note that all of the words in Table 5 that were found to occur with high frequency in social science texts (e.g., *political*, *ideological*, *societal*, *social*) also occurred with moderately high frequency in humanities texts. This suggests that a portion of the lack of fit in Figure 7 can be attributed to the difficulty associated with distinguishing acceptable social science texts from acceptable humanities texts. This issue will be considered further in additional planned research.

Evaluation via document classifications. The validity of the estimated acceptability probabilities was also evaluated by defining a *SourceFinder Acceptance Rule*, and then considering the extent to which the classification decisions generated under the specified rule agreed with the classification decisions provided by expert test developers.

The SourceFinder Acceptance Rule was defined as follows:

1. All documents with estimated acceptability probabilities at or above 0.5, for a particular assignment, were classified as acceptable for that assignment, and

2. All documents with estimated acceptability probabilities below 0.5, for a particular assignment, were classified as unacceptable for that assignment.

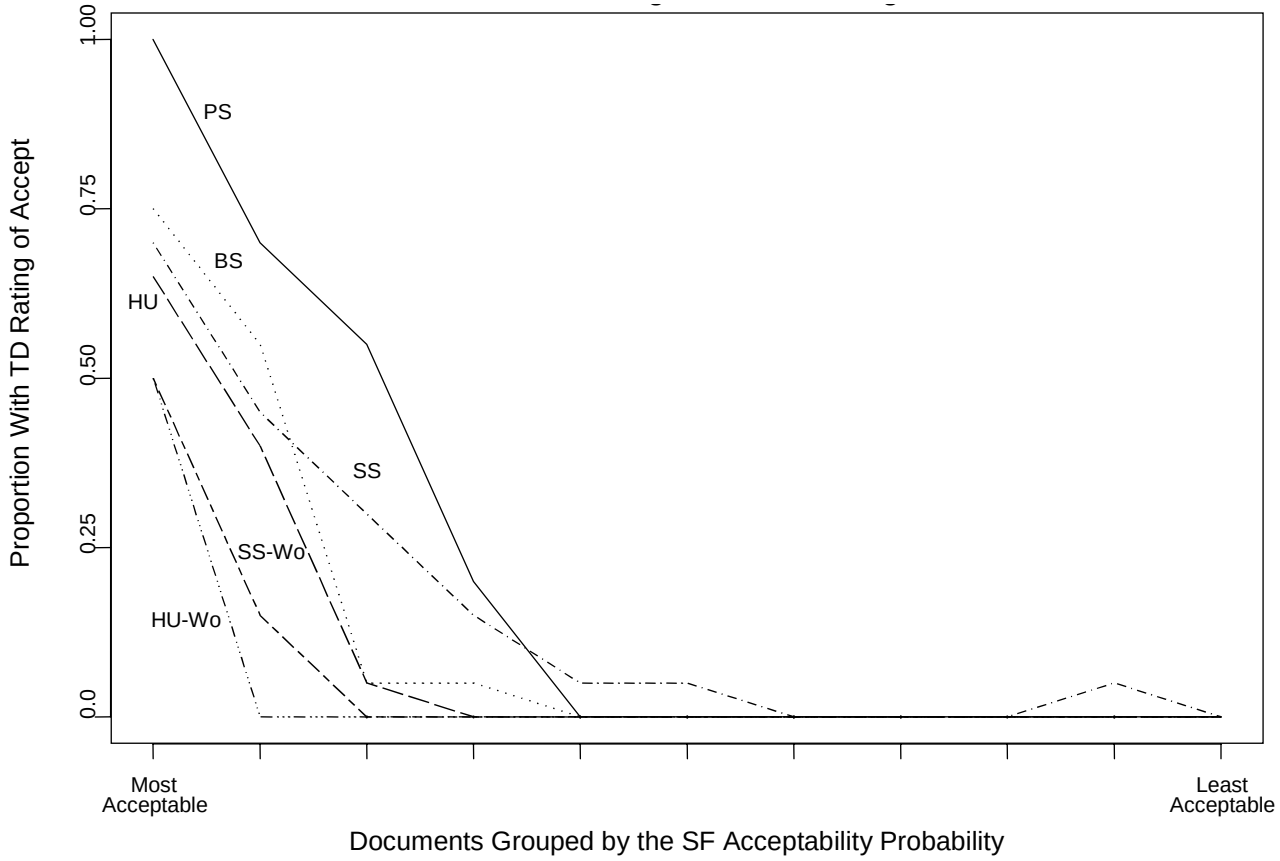


Figure 7. **Operating characteristic curves for six different passage-creation assignments.**

Note that this rule admits two different types of incorrect decisions: false positive decisions and false negative decisions. A false-positive decision occurs whenever SourceFinder accepts a document that is subsequently classified as unacceptable by all of the human raters. A false-negative decision occurs whenever SourceFinder rejects a document that is subsequently classified as acceptable by at least one of the human raters.

Table 8 shows the number and percent of correct and incorrect decisions obtained for the six assignments considered in this study. In considering these results, it is important to recall that, since a primary objective of the SourceFinder module is to help test developers find more acceptable sources in less time, false-positive decisions may be more costly than false-negative decisions. As is shown in Table 8, false-positive decisions were relatively rare, ranging from a low of 2% to a high of 5%.

Table 8

Classification Agreement Results for Specified Passage-Creation Assignments

Assignment/type of decision	Test developer classification	SourceFinder classification	Number	Percent	Total proportion correct
Physical Science					
Correct	Accept	Accept	40	18	
Correct	Reject	Reject	163	73	
False negative	Accept	Reject	9	4	
False positive	Reject	Accept	10	5	91%
Biological Science					
Correct	Accept	Accept	24	11	
Correct	Reject	Reject	185	83	
False negative	Accept	Reject	4	2	
False positive	Reject	Accept	9	4	94%
Social Science					
Correct	Accept	Accept	14	6	
Correct	Reject	Reject	182	82	
False negative	Accept	Reject	21	9	
False positive	Reject	Accept	5	2	88%
Humanities					
Correct	Accept	Accept	13	6	
Correct	Reject	Reject	193	87	
False negative	Accept	Reject	9	4	
False positive	Reject	Accept	7	3	93%
Social Science & women					
Correct	Accept	Accept	9	4	
Correct	Reject	Reject	207	93	
False negative	Accept	Reject	1	0	
False positive	Reject	Accept	5	2	97%
Humanities & women					
Correct	Accept	Accept	4	2	
Correct	Reject	Reject	205	92	
False negative	Accept	Reject	9	4	
False positive	Reject	Accept	4	2	94%

Additional reductions in the rate of false-positive decisions were achieved through implementation of the sensitivity, accessibility, and level of argumentation filters. These additional reductions are summarized in Table 9. The table confirms that the filters were successful at reducing the number of false-positive decisions while yielding only a slight increase in the rate of false-negative decisions. In considering these results, it is important to recall that the filters are designed to detect acceptability violations that are serious yet *rare*. As more and more of these rare violations are uncovered, it is likely that a more substantial reduction in the rate of false-positive decisions will be realized.

Table 9

Additional Reductions in the Rate of False-Positive and False-Negative Decisions Achieved Through the Filtering Process

Assignment	No. of false positive decisions			No. of false negative decisions		
	Before	After	Change	Before	After	Change
Physical Science	10	9	-1	9	9	0
Biological Science	9	7	-2	4	5	+1
Social Science–any	5	4	-1	1	3	+2
Humanities–any	7	3	-4	9	9	0
Social Science–women	5	5	0	1	1	0
Humanities–women	4	4	0	9	9	0
Total	40	32	-8	33	36	+3

Precision and recall. In many classification applications, performance is evaluated by considering the following conditional probabilities:

1. *P*(Human Classification = Accept | System Classification = Accept)
2. *P*(System Classification = Accept | Human Classification = Accept).

The first probability is referred to as the system's *precision*; the second probability is referred to as the system's *recall* (van Rijsbergen, 1979). In the application considered here, precision refers to the probability that a document that SourceFinder rates highly will also be rated highly by an expert test developer. When many of the documents that SourceFinder rates

highly are judged by expert test developers to be acceptable, precision will be high; otherwise it will be low. The second conditional probability focuses on SourceFinder's ability to locate the documents that the test developers tend to like; in other words, those that they rate as acceptable. When SourceFinder fails to locate a large proportion of the acceptable documents, recall will be low; otherwise it will be high.

Table 10 lists the precision and recall values estimated for the six assignments considered in this study. These values reflect the classification performance achieved through application of both the filtering process and the logistic regression process. For comparison purposes, the table also shows the precision and recall values obtained when the screening capability is turned off.

Table 10

The Precision and Recall of the Source Classification Process With Document Screening Turned On and With Document Screening Turned Off

Type of screening/ assignment	No. of training documents with TD ratings of accept	Precision *P*(TD = accept | SF = accept)	Recall *P*(SF = accept | TD = accept)
Screening = on[a]			
Physical Science	49	.82	.82
Biological Science	28	.77	.82
Social Science	35	.75	.34
Humanities	22	.81	.59
Social Science–women	10	.64	.90
Humanities–women	13	.50	.31
Screening = off [b]			
All assignments[c]	44	.32	1.00

[a] When document screening is turned on, a document is considered to have been accepted by SourceFinder only when the estimated acceptance probability for the specified assignment is at or above 0.5. [b] When document screening is not turned on, all retrieved documents are treated as having been accepted by SourceFinder. Note that, since no documents are screened out, the recall rate is necessarily 100%. [c] The 86 historical sources are not included here because the total number of candidate sources evaluated during the process of locating those sources is not known.

Table 10 shows that precision was quite high for the assignments that did not include a subcontent classification, and moderately high for the assignments that did include a subcontent classification. The practical significance of these results may be evaluated by comparing them to the null case; that is, the precision expected when the document screening capability is turned off (Screening = No). As is shown in Table 10, when the document screening capability is turned off, test developers can expect to find about 44 acceptable sources in each group of 136 documents examined. If each examined document is viewed as a document with a null classification of Accept, this translates into a null precision of about 32%. The precision of the newly developed screening capability may be evaluated by comparing its precision to this baseline figure. This comparison shows that all six of the estimated models resulted in significant gains in precision. For example, in the case of the Humanities-Women assignment, precision increased from about 32% to about 50%. Somewhat higher increases were achieved for each of the other assignments. For example, the Biological Sciences screening capability yielded an increase from 32% to 77%, and the Physical Sciences screening capability yielded an increase from 32% to 82%. These increases suggest that the current feature pool has succeeded in capturing important test-developer (TD) acceptability requirements, especially for the assignments that do not include subcontent restrictions.

Table 10 also lists the recall values estimated for each of the assignments. The table shows that high recall rates were obtained for some of the assignments (Physical Science, Biological Science, and Social Science-Women) but not for others. Since recall is low when important text characteristics are excluded from the estimated acceptability model, this suggests that additional feature development work is needed.

Cross-Validation

A question of interest is whether the precision increases reported in Table 10 will persist when the estimated models are applied to a new sample of documents drawn from the Source Database. Because the resources needed to obtain TD ratings for a new sample of candidate source documents were not available this question was investigated via a cross-validation analysis. The analysis was implemented as follows. First, the current sample of 222 candidate source documents was divided into estimation and validation subsets via a 74/26 split. That is, 74% of the documents were randomly assigned to the estimation dataset, while the remaining 26% were reserved for the cross-validation dataset. Second, the number of acceptable source

documents in each dataset, relative to each of the specified passage-creation assignments, was calculated. This calculation showed that only the Physical Sciences assignment yielded at least 10 acceptable sources in both the training and validation subsets. Because it was believed that a minimum of 10 acceptable sources would be needed to obtain interpretable results, the cross-validation analysis was implemented for the Physical Sciences passage-creation assignment only.

The analysis was implemented as follows. First, a new Physical Sciences prediction model was estimated from the documents in the training subset (n = 164). Second, this model was applied to each of the documents in the cross-validation subset (n = 58). Finally, separate performance summaries were prepared for the 164 documents in the training dataset, and for the 58 documents in the cross-validation dataset. These summaries document the performance of the logistic regression process only. The additional reductions in the rate of false-positive decisions achievable through application of the filtering process are not considered in these summaries.

Results are shown in Tables 11 and 12. Table 11 lists the number and percentage of correct and incorrect decisions observed in each dataset. Table 12 summarizes the effects on both precision and recall. Note that the strategy of validating the model on an independent set of documents did not result in a serious degradation of performance. This suggests that the logistic regression model estimated for the Physical Science assignment is likely to perform adequately when applied to new, previously unseen documents.

Conclusions

Both the current study, and previous research reported in Passonneau et al. (2002) confirmed that acceptable sources for GRE passage development are not abundant. In both studies, the acceptance rate for candidate source documents selected from appropriate scientific and literary journals was only about 32%. This low percentage of acceptable documents suggests that significant savings in the time needed to locate suitable source documents are possible.

A new, statistically based document-screening capability was incorporated into the SourceFinder architecture in June 2003. This new capability employs a combination of filtering and logistic regression techniques to assign a set of k acceptability probabilities to each candidate source document: one for each of k predefined passage-creation assignments. Test developers can use this capability to obtain lists of candidate source documents sorted from most acceptable to lease acceptable for specified source-finding assignments.

Table 11

Classification Results in the Training and Cross-Validation Data Sets for the Physical Sciences Logistic Regression Model

Type of sample/ type of decision	Test-developer classification	SourceFinder classification	Number	Percentage
Training sample (n = 164)				
Correct	Accept	Accept	34	21
Correct	Reject	Reject	123	75
False negative	Accept	Reject	1	1
False positive	Reject	Accept	6	4
Cross validation sample (n = 58)				
Correct	Accept	Accept	13	22
Correct	Reject	Reject	41	71
False negative	Accept	Reject	1	2
False positive	Reject	Accept	3	5

Table 12

The Precision and Recall of the Physical Science Acceptability Model in the Training and Cross-Validation Data Sets

Type of sample	No. of documents with TD ratings of accept	Precision P(TD =1 | SF >=0.5)	Recall P(SF >=0.5 | TD = 1)
Training (n = 164)	35	.85	.97
Cross-validation (n = 58)	14	.81	.93

This new capability was evaluated by comparing the proportion of acceptable documents obtained with the screening capability turned off to the proportion of acceptable documents obtained with the screening capability turned on. The evaluation confirmed that significant increases in the percent of acceptable documents located were achieved for each of the individual source-finding assignments considered. In particular, for the GRE Humanities-Women reading-

comprehension assignment, the acceptance rate achieved with the screening capability turned off was about 32%. When the screening capability was turned on, 50% of the retrieved documents were observed to be acceptable. A similar increase was observed for the Social Sciences-Women assignment: The rate of true accepts increased from 32% to 64%. Somewhat higher increases were observed for the four assignments that did *not* include a subcontent classification. In particular, the Social Sciences, Humanities, Biological Sciences, and Physical Sciences reading-comprehension assignments all yielded acceptance rates at or above 75%. These increases confirm that the new, statistically based document-screening capability documented in this paper represents a substantial improvement over the previous approach of no screening. The observed increases also suggest that the current feature pool has succeeded in capturing important TD acceptability requirements, and that the resulting capability can help test developers find more high-quality sources in less time.

Limitations

The single most important limitation of the current study was the lack of a large database of previously rated candidate source documents for use in model training and evaluation. This situation occurred because, in the past, information about the acceptability status of candidate source documents was not saved. Recent changes to the SourceFinder interface have been designed to ensure that this same limitation is not a factor in future research. In particular, the current interface has been updated to include a data-collection capability that allows test developers to efficiently record their judgments of source acceptability *at the time that those judgments are being made*. This new capability is designed to minimize interruptions to the creative process while simultaneously collecting high-quality data at minimum cost. The resulting database of expert judgments will be made available to researchers engaged in future development and evaluation studies.

Directions for Future Research

While the statistically based document-screening models documented in this report represent a clear improvement over the previous case of no screening, additional improvements are still needed. Additional planned analyses are discussed below.

Additional Feature-Development Work

The current study has confirmed that many of the aspects of text variation that contribute to test developers' judgments of source acceptability may be approximated by text features that can be automatically extracted by NLP tools. The low recall values obtained for some assignments suggest, however, that additional feature-development work is needed. Several modifications to the SourceFinder interface have been implemented to support this additional work. For example, a capability for capturing information about the individual words and phrases in actual sources that are indicative of an appropriate level of argumentation or of particular types of violations (e.g., language that is indicative of a serious violation of the GRE sensitivity standard) has been added. Authorized SourceFinder users now have the option of launching this tool whenever a particularly illustrative text is discovered during normal source processing. Once launched, the tool automatically copies highlighted words and phrases from the source document under consideration to a designated storage location in the Source database. It is expected that, over time, the words and phrases identified in this manner will help us to develop new features that will be even more predictive of the source-acceptability judgments made by expert test developers.

It is important to note that as more and more text features are developed for consideration in the models, the need for valid feature-selection and/or dimensionality-reduction techniques is likely to become more pressing. Thus, additional research aimed at evaluating alternative feature-selection and/or dimensionality-reduction techniques is also needed. Exploratory factor analyses have frequently been used to implement feature selection and dimensionality reduction in text-classification applications (see Biber, 1988; Biber et al., 2004; and Reppen, 2001). Biber et al. (2004, citing Ervin-Tripp, 1972) argued that, because many important dimensions of text variation are not well captured by individual linguistic features, investigation of such characteristics requires a focus on "constellations of co-occurring linguistic features" rather than on individual features. Factor analysis permits easy access to such *constellations* by allowing patterns of linguistic co-occurrence to be analyzed in terms of underlying *dimensions of variation* or *factors* that are identified quantitatively. An evaluation of alternative feature-selection and dimensionality-reduction techniques is in progress and will be reported in a future study.

Additional Validation Analyses

A problem encountered in the current investigation is that only a small sample of TD ratings was available for use in the analyses. This limitation suggests that additional research involving larger samples of TD ratings is needed. Also, it is important to recall that all of the documents considered in the current analyses were extracted from a set of scientific and literary journals that had been previously classified as highly appropriate for use in developing GRE passages and items, and that the Source Database has since been updated to include documents extracted from *30 additional* journals and magazines. Thus, the validity of the existing models, when applied to articles extracted from these lesser-known journals, has not yet been investigated, suggesting that a detailed investigation of model performance for articles extracted from the updated Source Database is also needed.

References

Bauer, M., & Jha, K. (1999). *A tool for finding source material for GRE reading comprehension items.* Princeton, NJ: ETS.

Biber, D. (1988). *Variation across speech and writing.* Cambridge, England: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., et al. (2004, January). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus* (TOEFL Monograph Series No. MS-25). Princeton, NJ: ETS.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Essex, England: Pearson Education Limited.

Brieman, L., Friedman, J. H., Olshen, R., & Stone, C. J. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth International Group.

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113-122). Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., et al. (1998). *Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays* (ETS RR-98-15). Princeton, NJ: ETS.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis.* Belmont, CA: Wadsworth International Group.

Ervin-Tripp, S. (1972). On sociolinguistic rules: Alternation and co-occurrence. In J. J Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics* (pp. 213-250). New York: Holt, Rinehart & Winston.

Harris, Z. S. (1968). *Mathematical structures of language.* New York: Wiley.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning.* New York: Springer-Verlag.

Jha, K. (2001). User manual for using the SourceFinder: Version 0.9 [Computer software manual]. Princeton, NJ: ETS.

Leydesdorff, L. (2005). *Similarity measures, author cocitation analysis, and information theory. Journal of the American society for Information Science and Technology, 56*(7), 769-772.

Lin, D. (1998). Automatic retrieval and clustering of similar words. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Montreal*, 898-904.

Passonneau, R., Hemat, L., Plante, J., & Sheehan, K. (2002). *Electronic sources as input to GRE reading comprehension item development: SourceFinder prototype evaluation.* (ETS RR-02-12). Princeton, NJ: ETS.

Reppen, R. (2001). Register variation in student and adult speech and writing. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 187-199). London: Longman.

Salton G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer.* Reading, MA: Addison-Wesley Publishing.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* New York: McGraw Hill.

Sheehan, K. M. (2003). Tree-based regression: A new tool for understanding cognitive skill requirements. In H. F. O'Neil & R. S. Perez (Eds.), *Technology applications in education: A learning view* (pp. 222-227). Mahwah, NJ: Lawrence Erlbaum Associates.

Sheehan, K. M., Kostin, I., & Futagi, Y. (2006). *Supporting efficient, evidence-centered item development for the GRE paragraph reading item type.* Manuscript in preparation.

Van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.), London: Butterworths.

Youmans, G. (1991). A new tool for discourse analysis: The vocabulary-management profile. *Language, 67*, 763-789.

Notes

[1] It is useful to note that this system diverges from the previous system in its handling of unacceptable documents. In particular, documents rated as having low acceptability probabilities are not discarded. Rather, these documents are retained for consideration by other source-finding applications.

[2] The GRE portion of the Source Database has since been updated to include over 90,000 source documents extracted from 60 different journals and magazines.

Appendix

Table A1

Document Ranks Induced by the GRE Biological Science Acceptability Model

Rank	SF	TD	Rank	SF	TD	Rank	SF	TD	Rank	SF	TD
1	0.9630	1	56	0.0108	0	111	0.0042	0	166	0.0026	0
2	0.9420	1	57	0.0092	0	112	0.0042	0	167	0.0026	0
3	0.9288	1	58	0.0088	0	113	0.0042	0	168	0.0025	0
4	0.9189	1	59	0.0088	0	114	0.0042	0	169	0.0025	0
5	0.8927	1	60	0.0087	0	115	0.0041	0	170	0.0025	0
6	0.8824	1	61	0.0086	0	116	0.0040	0	171	0.0025	0
7	0.8437	1	62	0.0086	0	117	0.0040	0	172	0.0025	0
8	0.8247	1	63	0.0080	0	118	0.0040	0	173	0.0025	0
9	0.8068	1	64	0.0079	0	119	0.0039	0	174	0.0025	0
10	0.7770	1	65	0.0079	0	120	0.0039	0	175	0.0024	0
11	0.7711	1	66	0.0078	0	121	0.0038	0	176	0.0024	0
12	0.7692	0	67	0.0078	1	122	0.0038	0	177	0.0024	0
13	0.7583	1	68	0.0076	0	123	0.0038	0	178	0.0024	0
14	0.7500	0	69	0.0075	0	124	0.0037	0	179	0.0023	0
15	0.7195	0	70	0.0075	0	125	0.0037	0	180	0.0023	0
16	0.6932	1	71	0.0075	0	126	0.0036	0	181	0.0023	0
17	0.6554	1	72	0.0073	0	127	0.0036	0	182	0.0023	0
18	0.6469	0	73	0.0072	0	128	0.0036	0	183	0.0023	0
19	0.6323	1	74	0.0068	0	129	0.0034	0	184	0.0022	0
20	0.6301	0	75	0.0068	0	130	0.0034	0	185	0.0022	0
21	0.6267	0	76	0.0067	0	131	0.0034	0	186	0.0021	0
22	0.6253	0	77	0.0066	0	132	0.0034	0	187	0.0021	0
23	0.6241	1	78	0.0065	0	133	0.0034	0	188	0.0021	0
24	0.6131	1	79	0.0065	0	134	0.0034	0	189	0.0021	0
25	0.5739	0	80	0.0064	0	135	0.0034	0	190	0.0021	0

(Table continues)

Table A1 (continued)

Rank	SF	TD	Rank	SF	TD	Rank	SF	TD	Rank	SF	TD
26	0.5720	1	81	0.0062	0	136	0.0034	0	191	0.0021	0
27	0.5556	1	82	0.0062	0	137	0.0033	0	192	0.0021	0
28	0.5536	1	83	0.0061	0	138	0.0033	0	193	0.0020	0
29	0.5466	1	84	0.0061	0	139	0.0033	0	194	0.0020	0
30	0.5359	1	85	0.0058	0	140	0.0033	0	195	0.0020	0
31	0.5270	1	86	0.0058	0	141	0.0033	0	196	0.0020	0
32	0.5211	0	87	0.0057	0	142	0.0033	0	197	0.0020	0
33	0.5011	1	88	0.0056	0	143	0.0033	0	198	0.0019	0
34	0.4868	0	89	0.0055	0	144	0.0032	0	199	0.0019	0
35	0.4837	0	90	0.0054	0	145	0.0032	0	200	0.0018	0
36	0.4827	0	91	0.0053	0	146	0.0032	0	201	0.0018	0
37	0.4775	0	92	0.0052	0	147	0.0032	0	202	0.0018	0
38	0.4528	1	93	0.0051	0	148	0.0031	0	203	0.0018	0
39	0.4206	1	94	0.0051	0	149	0.0031	0	204	0.0018	0
40	0.3741	0	95	0.0050	0	150	0.0031	0	205	0.0017	0
41	0.3388	1	96	0.0050	0	151	0.0031	0	206	0.0017	0
42	0.3011	0	97	0.0050	0	152	0.0030	0	207	0.0017	0
43	0.0783	0	98	0.0048	0	153	0.0030	0	208	0.0016	0
44	0.0582	0	99	0.0048	0	154	0.0029	0	209	0.0016	0
45	0.0466	0	100	0.0046	0	155	0.0029	0	210	0.0016	0
46	0.0287	0	101	0.0046	0	156	0.0028	0	211	0.0015	0
47	0.0246	0	102	0.0046	0	157	0.0028	0	212	0.0015	0
48	0.0244	0	103	0.0046	0	158	0.0028	0	213	0.0015	0
49	0.0223	0	104	0.0045	0	159	0.0028	0	214	0.0015	0
50	0.0147	0	105	0.0044	0	160	0.0028	0	215	0.0015	0
51	0.0129	0	106	0.0044	0	161	0.0027	0	216	0.0014	0
52	0.0117	0	107	0.0044	0	162	0.0027	0	217	0.0014	0
53	0.0116	0	108	0.0044	0	163	0.0027	0	218	0.0014	0
54	0.0112	0	109	0.0043	0	164	0.0027	0	219	0.0013	0
55	0.0110	0	110	0.0043	0	165	0.0026	0	220	0.0011	0
									221	0.0011	0
									222	0.0011	0

Table A2

Document Ranks Induced by the GRE Social Science Acceptability Model

Rank	SF	TD	Rank	SF	TD	Rank	SF	TD	Rank	SF	TD
1	0.9859	0	56	0.1934	0	111	0.0657	0	166	0.0408	0
2	0.9621	1	57	0.1923	0	112	0.0650	0	167	0.0405	0
3	0.9298	1	58	0.1911	0	113	0.0645	0	168	0.0403	0
4	0.9224	1	59	0.1874	0	114	0.0644	0	169	0.0392	0
5	0.8994	1	60	0.1838	1	115	0.0643	0	170	0.0386	0
6	0.8840	1	61	0.1822	1	116	0.0629	1	171	0.0382	0
7	0.8502	1	62	0.1788	0	117	0.0628	0	172	0.0379	0
8	0.8497	1	63	0.1753	0	118	0.0627	0	173	0.0360	0
9	0.8319	0	64	0.1743	0	119	0.0620	0	174	0.0359	0
10	0.8202	0	65	0.1439	1	120	0.0617	0	175	0.0359	0
11	0.7655	1	66	0.1370	0	121	0.0599	0	176	0.0349	0
12	0.7324	0	67	0.1337	0	122	0.0598	0	177	0.0347	0
13	0.6486	1	68	0.1331	0	123	0.0598	0	178	0.0342	0
14	0.6178	1	69	0.1261	0	124	0.0594	0	179	0.0312	0
15	0.5825	1	70	0.1240	0	125	0.0592	0	180	0.0305	0
16	0.5690	1	71	0.1215	0	126	0.0586	0	181	0.0294	0
17	0.5479	0	72	0.1205	0	127	0.0585	0	182	0.0293	0
18	0.5426	1	73	0.1200	0	128	0.0579	0	183	0.0292	0
19	0.5093	1	74	0.1140	1	129	0.0574	0	184	0.0289	0
20	0.4648	0	75	0.1088	0	130	0.0574	0	185	0.0282	0
21	0.4530	1	76	0.1087	0	131	0.0570	0	186	0.0281	0
22	0.4399	1	77	0.1030	0	132	0.0569	0	187	0.0269	0
23	0.4185	1	78	0.1019	0	133	0.0566	0	188	0.0266	0
24	0.4121	0	79	0.0999	0	134	0.0564	0	189	0.0264	0
25	0.3914	1	80	0.0965	0	135	0.0558	0	190	0.0262	0

(Table continues)

Table A2 (continued)

Rank	SF	TD	Rank	SF	TD	Rank	SF	TD	Rank	SF	TD
26	0.3886	0	81	0.0927	0	136	0.0556	0	191	0.0248	0
27	0.3610	0	82	0.0920	0	137	0.0554	0	192	0.0225	0
28	0.3522	0	83	0.0919	0	138	0.0550	0	193	0.0217	0
29	0.3351	0	84	0.0913	0	139	0.0546	0	194	0.0209	1
30	0.3262	0	85	0.0912	0	140	0.0546	0	195	0.0171	0
31	0.2985	0	86	0.0884	0	141	0.0545	0	196	0.0167	0
32	0.2932	1	87	0.0873	0	142	0.0533	0	197	0.0153	0
33	0.2883	0	88	0.0868	0	143	0.0526	0	198	0.0152	0
34	0.2873	0	89	0.0851	0	144	0.0526	0	199	0.0152	0
35	0.2863	0	90	0.0842	0	145	0.0521	0	200	0.0136	0
36	0.2828	0	91	0.0816	0	146	0.0509	0	201	0.0075	0
37	0.2825	1	92	0.0801	0	147	0.0508	0	202	0.0063	0
38	0.2821	1	93	0.0779	0	148	0.0498	0	203	0.0060	0
39	0.2798	1	94	0.0768	0	149	0.0492	0	204	0.0052	0
40	0.2647	1	95	0.0766	0	150	0.0480	0	205	0.0049	0
41	0.2554	1	96	0.0762	0	151	0.0473	0	206	0.0040	0
42	0.2545	0	97	0.0756	0	152	0.0471	0	207	0.0037	0
43	0.2479	0	98	0.0747	1	153	0.0469	0	208	0.0034	0
44	0.2457	1	99	0.0745	0	154	0.0462	0	209	0.0033	0
45	0.2403	0	100	0.0738	0	155	0.0461	0	210	0.0027	0
46	0.2275	1	101	0.0724	0	156	0.0456	0	211	0.0022	0
47	0.2254	1	102	0.0722	0	157	0.0449	0	212	0.0018	0
48	0.2253	0	103	0.0689	0	158	0.0440	0	213	0.0018	0
49	0.2244	0	104	0.0681	0	159	0.0439	0	214	0.0017	0
50	0.2207	0	105	0.0677	0	160	0.0435	0	215	0.0008	0
51	0.2205	0	106	0.0676	0	161	0.0429	0	216	0.0005	0
52	0.2197	0	107	0.0675	0	162	0.0429	0	217	0.0004	0
53	0.2154	1	108	0.0670	0	163	0.0422	0	218	0.0003	0
54	0.2136	0	109	0.0669	0	164	0.0411	0	219	0.0002	0
55	0.2013	0	110	0.0664	0	165	0.0408	0	220	0.0000	0
									221	0.0000	0
									222	0.0000	0

Table A3

Document Ranks Induced by the GRE Humanities Acceptability Model

Rank	SF	TD	Rank	SF	TD	Rank	SF	TD	Rank	SF	TD
1	0.9125	1	56	0.0581	0	111	0.0048	0	166	0.0000	0
2	0.9029	0	57	0.0567	0	112	0.0046	0	167	0.0000	0
3	0.8518	1	58	0.0554	0	113	0.0045	0	168	0.0000	0
4	0.8115	1	59	0.0488	0	114	0.0045	0	169	0.0000	0
5	0.7554	1	60	0.0478	0	115	0.0038	0	170	0.0000	0
6	0.6988	0	61	0.0402	0	116	0.0035	0	171	0.0000	0
7	0.6648	0	62	0.0387	0	117	0.0035	0	172	0.0000	0
8	0.6335	1	63	0.0353	0	118	0.0030	0	173	0.0000	0
9	0.5956	0	64	0.0336	0	119	0.0028	0	174	0.0000	0
10	0.5921	1	65	0.0325	0	120	0.0026	0	175	0.0000	0
11	0.5907	0	66	0.0305	0	121	0.0024	0	176	0.0000	0
12	0.5782	0	67	0.0283	0	122	0.0022	0	177	0.0000	0
13	0.5436	0	68	0.0270	0	123	0.0022	0	178	0.0000	0
14	0.5402	1	69	0.0230	0	124	0.0022	0	179	0.0000	0
15	0.5352	0	70	0.0225	0	125	0.0018	0	180	0.0000	0
16	0.5295	1	71	0.0219	0	126	0.0015	0	181	0.0000	0
17	0.5149	0	72	0.0218	0	127	0.0013	0	182	0.0000	0
18	0.4871	1	73	0.0218	0	128	0.0013	0	183	0.0000	0
19	0.4433	1	74	0.0211	0	129	0.0013	0	184	0.0000	0
20	0.4287	1	75	0.0206	0	130	0.0012	0	185	0.0000	0
21	0.4200	0	76	0.0202	0	131	0.0012	0	186	0.0000	0
22	0.4176	1	77	0.0142	0	132	0.0012	0	187	0.0000	0
23	0.4114	1	78	0.0140	0	133	0.0012	0	188	0.0000	0
24	0.4041	0	79	0.0137	0	134	0.0011	0	189	0.0000	0
25	0.3760	1	80	0.0137	0	135	0.0011	0	190	0.0000	0

(Table continues)

Table A3 (continued)

Rank	SF	TD	Rank	SF	TD	Rank	SF	TD	Rank	SF	TD
26	0.3432	0	81	0.0135	0	136	0.0009	0	191	0.0000	0
27	0.3075	1	82	0.0134	0	137	0.0009	0	192	0.0000	0
28	0.2923	1	83	0.0132	0	138	0.0009	0	193	0.0000	0
29	0.2520	0	84	0.0129	0	139	0.0009	0	194	0.0000	0
30	0.2201	1	85	0.0127	0	140	0.0008	0	195	0.0000	0
31	0.2144	0	86	0.0114	0	141	0.0008	0	196	0.0000	0
32	0.1863	1	87	0.0112	0	142	0.0008	0	197	0.0000	0
33	0.1802	0	88	0.0104	0	143	0.0008	0	198	0.0000	0
34	0.1784	0	89	0.0098	0	144	0.0006	0	199	0.0000	0
35	0.1772	1	90	0.0097	0	145	0.0006	0	200	0.0000	0
36	0.1442	0	91	0.0096	0	146	0.0006	0	201	0.0000	0
37	0.1353	0	92	0.0095	0	147	0.0004	0	202	0.0000	0
38	0.1295	0	93	0.0090	0	148	0.0004	0	203	0.0000	0
39	0.1240	1	94	0.0090	0	149	0.0003	0	204	0.0000	0
40	0.1181	0	95	0.0084	0	150	0.0002	0	205	0.0000	0
41	0.1156	0	96	0.0081	0	151	0.0002	0	206	0.0000	0
42	0.1125	0	97	0.0076	0	152	0.0002	0	207	0.0000	0
43	0.1053	0	98	0.0067	0	153	0.0002	0	208	0.0000	0
44	0.1047	0	99	0.0064	0	154	0.0001	0	209	0.0000	0
45	0.1001	0	100	0.0063	0	155	0.0001	0	210	0.0000	0
46	0.0905	0	101	0.0060	0	156	0.0001	0	211	0.0000	0
47	0.0879	0	102	0.0060	0	157	0.0001	0	212	0.0000	0
48	0.0852	0	103	0.0060	0	158	0.0000	0	213	0.0000	0
49	0.0809	0	104	0.0055	0	159	0.0000	0	214	0.0000	0
50	0.0753	0	105	0.0054	0	160	0.0000	0	215	0.0000	0
51	0.0731	0	106	0.0052	0	161	0.0000	0	216	0.0000	0
52	0.0615	0	107	0.0051	0	162	0.0000	0	217	0.0000	0
53	0.0606	0	108	0.0050	0	163	0.0000	0	218	0.0000	0
54	0.0595	0	109	0.0049	0	164	0.0000	0	219	0.0000	0
55	0.0592	0	110	0.0049	0	165	0.0000	0	220	0.0000	0
									221	0.0000	0
									222	0.0000	0

Table A4

Document Ranks Induced by the GRE Social Science-Women Acceptability Model

Rank	SF	TD	Rank	SF	TD	Rank	SF	TD	Rank	SF	TD
1	0.9583	1	56	0.0085	0	111	0.0012	0	166	0.0005	0
2	0.9356	0	57	0.0082	0	112	0.0012	0	167	0.0005	0
3	0.8399	1	58	0.0075	0	113	0.0012	0	168	0.0005	0
4	0.8237	1	59	0.0068	0	114	0.0012	0	169	0.0005	0
5	0.8196	0	60	0.0068	0	115	0.0011	0	170	0.0005	0
6	0.7655	1	61	0.0064	0	116	0.0011	0	171	0.0005	0
7	0.7324	0	62	0.0063	0	117	0.0011	0	172	0.0005	0
8	0.6957	0	63	0.0060	0	118	0.0011	0	173	0.0005	0
9	0.6616	1	64	0.0060	0	119	0.0011	0	174	0.0005	0
10	0.6431	1	65	0.0055	0	120	0.0010	0	175	0.0005	0
11	0.5823	1	66	0.0051	0	121	0.0010	0	176	0.0005	0
12	0.5588	1	67	0.0049	0	122	0.0010	0	177	0.0004	0
13	0.5478	0	68	0.0048	0	123	0.0010	0	178	0.0004	0
14	0.5357	1	69	0.0047	0	124	0.0010	0	179	0.0004	0
15	0.3350	0	70	0.0046	0	125	0.0009	0	180	0.0004	0
16	0.2985	0	71	0.0045	0	126	0.0008	0	181	0.0004	0
17	0.2847	1	72	0.0043	0	127	0.0008	0	182	0.0004	0
18	0.2668	0	73	0.0040	0	128	0.0008	0	183	0.0004	0
19	0.2275	0	74	0.0040	0	129	0.0008	0	184	0.0003	0
20	0.1420	0	75	0.0038	0	130	0.0008	0	185	0.0003	0
21	0.1153	0	76	0.0038	0	131	0.0008	0	186	0.0003	0
22	0.0870	0	77	0.0038	0	132	0.0008	0	187	0.0003	0
23	0.0813	0	78	0.0037	0	133	0.0008	0	188	0.0003	0
24	0.0809	0	79	0.0036	0	134	0.0007	0	189	0.0003	0
25	0.0732	0	80	0.0036	0	135	0.0007	0	190	0.0003	0

(Table continues)

Table A4 (continued)

Rank	SF	TD	Rank	SF	TD	Rank	SF	TD	Rank	SF	TD
26	0.0680	0	81	0.0033	0	136	0.0007	0	191	0.0003	0
27	0.0669	0	82	0.0033	0	137	0.0007	0	192	0.0003	0
28	0.0469	0	83	0.0032	0	138	0.0007	0	193	0.0003	0
29	0.0447	0	84	0.0032	0	139	0.0007	0	194	0.0003	0
30	0.0368	0	85	0.0031	0	140	0.0007	0	195	0.0003	0
31	0.0353	0	86	0.0030	0	141	0.0007	0	196	0.0003	0
32	0.0336	0	87	0.0030	0	142	0.0007	0	197	0.0003	0
33	0.0218	0	88	0.0029	0	143	0.0007	0	198	0.0003	0
34	0.0208	0	89	0.0027	0	144	0.0007	0	199	0.0002	0
35	0.0207	0	90	0.0026	0	145	0.0007	0	200	0.0002	0
36	0.0196	0	91	0.0026	0	146	0.0007	0	201	0.0002	0
37	0.0174	0	92	0.0026	0	147	0.0007	0	202	0.0002	0
38	0.0164	0	93	0.0026	0	148	0.0007	0	203	0.0002	0
39	0.0160	0	94	0.0025	0	149	0.0007	0	204	0.0002	0
40	0.0160	0	95	0.0023	0	150	0.0007	0	205	0.0002	0
41	0.0157	0	96	0.0023	0	151	0.0007	0	206	0.0002	0
42	0.0156	0	97	0.0022	0	152	0.0006	0	207	0.0002	0
43	0.0147	0	98	0.0021	0	153	0.0006	0	208	0.0002	0
44	0.0145	0	99	0.0020	0	154	0.0006	0	209	0.0002	0
45	0.0141	0	100	0.0018	0	155	0.0006	0	210	0.0002	0
46	0.0136	0	101	0.0017	0	156	0.0006	0	211	0.0002	0
47	0.0121	0	102	0.0016	0	157	0.0006	0	212	0.0002	0
48	0.0119	0	103	0.0016	0	158	0.0006	0	213	0.0001	0
49	0.0119	0	104	0.0015	0	159	0.0006	0	214	0.0001	0
50	0.0099	0	105	0.0014	0	160	0.0006	0	215	0.0001	0
51	0.0097	0	106	0.0014	0	161	0.0006	0	216	0.0001	0
52	0.0093	0	107	0.0014	0	162	0.0006	0	217	0.0000	0
53	0.0089	0	108	0.0014	0	163	0.0005	0	218	0.0000	0
54	0.0089	0	109	0.0014	0	164	0.0005	0	219	0.0000	0
55	0.0089	0	110	0.0013	0	165	0.0005	0	220	0.0000	0
									221	0.0000	0
									222	0.0000	0

Table A5

Document Ranks Induced by the GRE Humanities-Women Acceptability Model

Rank	SF	TD	Rank	SF	TD	Rank	SF	TD	Rank	SF	TD
1	0.8511	1	56	0.0042	0	111	0.0001	0	166	0.0000	0
2	0.7447	1	57	0.0037	0	112	0.0001	0	167	0.0000	0
3	0.5781	0	58	0.0032	0	113	0.0001	0	168	0.0000	0
4	0.5381	1	59	0.0022	0	114	0.0001	0	169	0.0000	0
5	0.5351	0	60	0.0022	0	115	0.0001	0	170	0.0000	0
6	0.5295	1	61	0.0022	0	116	0.0001	0	171	0.0000	0
7	0.5275	0	62	0.0019	0	117	0.0001	0	172	0.0000	0
8	0.4658	0	63	0.0017	0	118	0.0001	0	173	0.0000	0
9	0.4175	1	64	0.0016	0	119	0.0001	0	174	0.0000	0
10	0.4098	1	65	0.0015	0	120	0.0000	0	175	0.0000	0
11	0.3961	1	66	0.0013	0	121	0.0000	0	176	0.0000	0
12	0.3156	0	67	0.0013	0	122	0.0000	0	177	0.0000	0
13	0.2733	1	68	0.0012	0	123	0.0000	0	178	0.0000	0
14	0.2035	0	69	0.0012	0	124	0.0000	0	179	0.0000	0
15	0.1905	1	70	0.0009	0	125	0.0000	0	180	0.0000	0
16	0.1854	0	71	0.0008	0	126	0.0000	0	181	0.0000	0
17	0.1780	0	72	0.0008	0	127	0.0000	0	182	0.0000	0
18	0.1643	1	73	0.0008	0	128	0.0000	0	183	0.0000	0
19	0.1348	0	74	0.0008	0	129	0.0000	0	184	0.0000	0
20	0.1235	1	75	0.0007	0	130	0.0000	0	185	0.0000	0
21	0.1175	0	76	0.0007	0	131	0.0000	0	186	0.0000	0
22	0.1150	0	77	0.0007	0	132	0.0000	0	187	0.0000	0
23	0.1059	0	78	0.0006	0	133	0.0000	0	188	0.0000	0
24	0.1049	0	79	0.0006	0	134	0.0000	0	189	0.0000	0
25	0.1039	0	80	0.0005	0	135	0.0000	0	190	0.0000	0

(Table continues)

Table A5 (continued)

Rank	SF	TD	Rank	SF	TD	Rank	SF	TD	Rank	SF	TD
26	0.0853	0	81	0.0005	0	136	0.0000	0	191	0.0000	0
27	0.0847	0	82	0.0005	0	137	0.0000	0	192	0.0000	0
28	0.0806	0	83	0.0005	0	138	0.0000	0	193	0.0000	0
29	0.0753	0	84	0.0005	0	139	0.0000	0	194	0.0000	0
30	0.0718	0	85	0.0004	0	140	0.0000	0	195	0.0000	0
31	0.0525	1	86	0.0004	0	141	0.0000	0	196	0.0000	0
32	0.0524	0	87	0.0004	0	142	0.0000	0	197	0.0000	0
33	0.0484	0	88	0.0004	0	143	0.0000	0	198	0.0000	0
34	0.0481	0	89	0.0003	0	144	0.0000	0	199	0.0000	0
35	0.0365	0	90	0.0003	0	145	0.0000	0	200	0.0000	0
36	0.0362	1	91	0.0003	0	146	0.0000	0	201	0.0000	0
37	0.0319	0	92	0.0003	0	147	0.0000	0	202	0.0000	0
38	0.0313	0	93	0.0002	0	148	0.0000	0	203	0.0000	0
39	0.0296	0	94	0.0002	0	149	0.0000	0	204	0.0000	0
40	0.0292	0	95	0.0002	0	150	0.0000	0	205	0.0000	0
41	0.0267	0	96	0.0002	0	151	0.0000	0	206	0.0000	0
42	0.0240	0	97	0.0002	0	152	0.0000	0	207	0.0000	0
43	0.0224	0	98	0.0002	0	153	0.0000	0	208	0.0000	0
44	0.0198	0	99	0.0001	0	154	0.0000	0	209	0.0000	0
45	0.0172	0	100	0.0001	0	155	0.0000	0	210	0.0000	0
46	0.0153	0	101	0.0001	0	156	0.0000	0	211	0.0000	0
47	0.0149	0	102	0.0001	0	157	0.0000	0	212	0.0000	0
48	0.0104	0	103	0.0001	0	158	0.0000	0	213	0.0000	0
49	0.0076	0	104	0.0001	0	159	0.0000	0	214	0.0000	0
50	0.0073	0	105	0.0001	0	160	0.0000	0	215	0.0000	0
51	0.0072	0	106	0.0001	0	161	0.0000	0	216	0.0000	0
52	0.0067	0	107	0.0001	0	162	0.0000	0	217	0.0000	0
53	0.0058	0	108	0.0001	0	163	0.0000	0	218	0.0000	0
54	0.0055	0	109	0.0001	0	164	0.0000	0	219	0.0000	0
55	0.0047	0	110	0.0001	0	165	0.0000	0	220	0.0000	0
									221	0.0000	0
									222	0.0000	0