Electronic Sources as Input to GRE® Reading Comprehension Item Development: SourceFinder Prototype Evaluation

Rebecca Passonneau
Lisa Hemat
Jan Plante
Kathleen M. Sheehan

June 2002

GRE Board Professional Report No. 99-18P
ETS Research Report 02-12
Electronic Sources as Input to
GRE® Reading Comprehension Item Development:
SourceFinder Prototype Evaluation

Rebecca Passonneau
Columbia University, New York

Lisa Hemat, Jan Plante, and Kathleen M. Sheehan
Educational Testing Service

GRE Board Report No. 99-18P

June 2002

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board and Educational Testing Service.

Educational Testing Service, Princeton, NJ 08541
Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board reports do not necessarily represent official Graduate Record Examinations Board position or policy.

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logos, GRADUATE RECORD EXAMINATIONS, and GRE are registered trademarks of Educational Testing Service. GMAT and GRADUATE MANAGEMENT ADMISSION TEST are registered trademarks of the Graduate Management Admission Council.

Educational Testing Service
Princeton, NJ 08541

Copyright © 2002 by Educational Testing Service. All rights reserved.
Abstract

This evaluation study compares the performance of a prototype tool called SourceFinder against the performance of highly trained human test developers. SourceFinder — a specialized search engine developed to locate source material for Graduate Record Examinations (GRE®) reading comprehension passages — employs a variety of shallow linguistic features to model the search criteria employed by expert test developers, to automate the source selection process, and to reduce source-processing time. The current evaluation provides detailed information about the aspects of source variation that are not well modeled by the current prototype. Approaches for enhancing performance in identified areas are discussed. The present study also provides a more explicit description of the source selection task, and a rich data set for developing a less subjective, more explicit definition of the types of documents preferred by test developers.

Key words: GRE, SourceFinder, reading comprehension stimuli, test development, natural language processing, NLP
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>GRE® Reading Comprehension Passages</td>
<td>1</td>
</tr>
<tr>
<td><strong>The SourceFinder Prototype System</strong></td>
<td>3</td>
</tr>
<tr>
<td>System Overview</td>
<td>3</td>
</tr>
<tr>
<td>SourceFinder System Processing</td>
<td>5</td>
</tr>
<tr>
<td>1999 Preliminary Evaluation</td>
<td>7</td>
</tr>
<tr>
<td><strong>Method</strong></td>
<td>7</td>
</tr>
<tr>
<td>Modifications to the SourceFinder Prototype</td>
<td>7</td>
</tr>
<tr>
<td><strong>Corpus Creation</strong></td>
<td>8</td>
</tr>
<tr>
<td>Document Selection</td>
<td>10</td>
</tr>
<tr>
<td>Allocation of Documents to Raters</td>
<td>11</td>
</tr>
<tr>
<td>Rater Instructions</td>
<td>12</td>
</tr>
<tr>
<td><strong>Results</strong></td>
<td>13</td>
</tr>
<tr>
<td>Corpus Creation</td>
<td>13</td>
</tr>
<tr>
<td>Evaluation of SourceFinder Against Human Judgments</td>
<td>14</td>
</tr>
<tr>
<td>Analysis of Expert Acceptability Judgments</td>
<td>14</td>
</tr>
<tr>
<td>Precision and Recall of SourceFinder’s Classification Algorithm</td>
<td>20</td>
</tr>
<tr>
<td>Rater Comments</td>
<td>23</td>
</tr>
<tr>
<td>A Simulation of Serial Filtering</td>
<td>23</td>
</tr>
<tr>
<td>Evaluation of Individual SourceFinder Features</td>
<td>26</td>
</tr>
<tr>
<td>First-Person Pronouns as a Genre Filter Enhancement</td>
<td>26</td>
</tr>
<tr>
<td>Document Normalization as a General Filter Enhancement</td>
<td>28</td>
</tr>
<tr>
<td>Analyses of Additional SourceFinder Features</td>
<td>30</td>
</tr>
<tr>
<td>Analysis of Source-to-Passage Feature Variation</td>
<td>32</td>
</tr>
<tr>
<td>SourceFinder Interface Improvements</td>
<td>33</td>
</tr>
</tbody>
</table>
List of Tables

Table 1. Two Sets of Threshold Values for SourceFinder .......................................................... 9
Table 2. Design for Allocating Documents to Raters In the Humanities Content Area .......... 11
Table 3. Summary of Classification Results Obtained During Corpus Creation ................. 14
Table 4. Sample Free-Form Comments and Subsequent Classifications ............................... 15
Table 5. Source Acceptability Ratings, by Rater, for 100 SourceFinder Classifications ....... 16
Table 6. Raters’ Joint Classification Results for 80 SourceFinder “Accepts” ............... 17
Table 7. Consensus Judgments of Four Expert Human Raters by Content Area ............. 19
Table 8. Individual Agreement With Consensus Judgments of Source Acceptability ........ 20
Table 9. Individual Agreement With Consensus Judgments of Source Acceptability by Rater 20
Table 10. Classification Results With Loose Acceptance Criteria ........................................ 21
Table 11. Classification Results With Tight Acceptance Criteria .......................................... 22
Table 12. A Classification System for Source Variation ......................................................... 24
Table 13. Sample Rater Comments by Category ....................................................... 25
Table 14. A Simulation of Serial Filtering by Content Area ................................................. 26
Table 15. Potential of First-Person Pronoun “I” as an Additional SourceFinder Filter ...... 28
Table 16. Feature Values of Full Text Versus Normalized Text ............................................. 30
Table 17. Correlation Results for SourceFinder’s UniqueWordperWord Feature .......... 31
Table 18. Correlation Results for SourceFinder’s KeyWordperWord Feature .......... 32
Table 19. Feature Comparison of Source Documents and Corresponding GRE Passages .... 33
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.</td>
<td>SourceFinder Database Screen, First Version</td>
<td>3</td>
</tr>
<tr>
<td>Figure 2.</td>
<td>SourceFinder User Interface, First Version</td>
<td>4</td>
</tr>
<tr>
<td>Figure 3.</td>
<td>A High-Level Overview of the SourceFinder System</td>
<td>5</td>
</tr>
<tr>
<td>Figure 4.</td>
<td>Decision-Process Graphic</td>
<td>13</td>
</tr>
<tr>
<td>Figure 5.</td>
<td>New SourceFinder User Interface: Search</td>
<td>37</td>
</tr>
<tr>
<td>Figure 6.</td>
<td>New SourceFinder User Interface: Browse Sources</td>
<td>38</td>
</tr>
<tr>
<td>Figure 7.</td>
<td>New SourceFinder User Interface: Review Entire Document</td>
<td>38</td>
</tr>
<tr>
<td>Figure 8.</td>
<td>New SourceFinder User Interface: Capture Test Developers’ Judgments</td>
<td>39</td>
</tr>
</tbody>
</table>
Introduction

SourceFinder was developed in 1999 to enable Educational Testing Service (ETS®) test developers to find reading comprehension stimuli for Graduate Record Examinations (GRE®) tests on the Web (Bauer & Jha, 1999; Jha, 2001). Development of the SourceFinder prototype was motivated by a desire to (a) reduce the time needed to locate acceptable sources; (b) expand the pool of documents available to test developers by integrating access to Web documents within the source location process; (c) regularize the process by moving away from a pencil-and-paper regimen to an electronic environment; (d) support the review process by allowing item writers to electronically record their reviews of sources to improve the management of sources; and (e) support the training of new item writers by accumulating a corpus of rejected and accepted sources with associated reviews. This early work, conducted using ETS Capability funds, resulted in a prototype version of a Web-crawler search engine.

The goals of the current research were to evaluate the existing SourceFinder prototype as a stand-alone tool, define more precisely the difficulty and nature of the test developers’ source finding task, and derive recommendations for further SourceFinder research and development based on a clearer understanding of the types of judgements made by highly trained humans. The project objectives therefore included:

1. Perform an evaluation of the existing SourceFinder prototype to determine its effectiveness for locating suitable GRE reading comprehension sources by collecting expert judgments from GRE test developers and comparing them to SourceFinder’s accept/reject classifications.

2. Identify additional functionality that will enable GRE test developers to interact more effectively with the SourceFinder prototype system and to shift from a PC-based system to client-server architecture.

3. Create a corpus of source documents to use in subsequent software development activity, and identify key research areas that will likely lead to additional improvements to SourceFinder.

GRE Reading Comprehension Passages

The reading comprehension passages that appear on the verbal section of the GRE General Test are adapted from previously published sources, such as journal articles. Results of interviews with test developers suggested that certain journals — such as those found in the
According to test developers, a source must be sufficiently challenging, but not obscure, to be usable; source text must meet stringent requirements on content, density, clarity, style, and length. The following excerpt from an internal GRE training manual describes the types of constraints that must be satisfied:

Reading comprehension passages generally rely on some kind of “tension” to give them difficulty. A discussion or argument that develops in a linear fashion toward a conclusion is usually too simple to support challenging items. The requisite tension can take various forms: a conflict between different or opposing arguments about something, an unexpected finding that challenges previously established ways of understanding something, a disagreement about what evidence is relevant to the solution of a particular problem, etc. But disagreement or contrast or opposition is not in itself enough to provide the tension: a passage that merely states or describes opposing positions, without discussing in much detail the reasoning behind them, will not support very many or very complex items. For example, instead of saying that Brown believed X and Smith believed Y, a passage might tell us that Brown, heavily influenced by such-and-such school of thought, tended to assume A, and therefore believed X. Smith, on the other hand, had access to information that Brown did not have, namely B and C, and this information led Smith to conclude Y. However, Smith interpreted B in such a way that it seemed logical to conclude Y, when in fact this interpretation was ill founded. The more interdependent details, the more logical twists and turns, the better. Naturally, a passage containing this kind of density can only be created from a source that provides the requisite information, and such sources are not easy to find. (p.4)

As the above excerpt suggests, the process of locating suitable source material for GRE reading comprehension passages is both difficult and time consuming. For every document that is accepted, a much larger number of potentially useful documents will likely have been retrieved, evaluated, and rejected by test developers. Also, there is a significant tradeoff between the characteristics of the source document and the amount of effort involved in crafting a usable
passage. That is, given a better “fit” of the source document, there will be significantly less editing involved in crafting a passage. Communications with test developers indicated that they sometimes resort to spending more time on the editing phase when the fit is suboptimal, rather than search for a new source. Developing skill with this task and managing the tradeoff between time spent locating a source and time spent crafting a passage from the source requires training and experience.

**The SourceFinder Prototype System**

**System Overview**

The original SourceFinder prototype is helpful in two essential ways. First, as Figure 1 shows, it automatically gathers potentially useful electronic documents and stores them in a searchable central, online database. Second, using the graphical user interface shown in Figure 2, test developers can selectively search, retrieve, review, and annotate documents in this database. These two capabilities were designed to reduce the time needed for retrieving, evaluating, and classifying potential source material.

![SourceFinder database screen, first version.](image)

**Figure 1.** SourceFinder database screen, first version.
Note: Tests Developers search by (1) Web domain, (2) program, and (3) source length.

Figure 2. SourceFinder user interface, first version.

As Figure 3 illustrates, the SourceFinder system is made up of four main components:

- **Text Gatherer** — When set to run as a Web crawler, this component automatically retrieves Web pages (i.e., candidate source material) from the Internet by starting at a user-specified site and branching to linked pages and sites.

- **Text Extractor** (renamed Text Converter) — This component converts retrieved Web pages from their original format (i.e., .html or .pdf) to plain (i.e., ASCII) text and then segments the texts by paragraph. Individual texts that meet minimum length requirements are retained for further analysis.

- **Text Selector** — Candidate texts are characterized by this component using a variety of shallow linguistic filters, such as word frequency, sentence length, and discourse cue features. Documents that meet specified filtering criteria are stored in the Candidate Source Database for later retrieval. (Appendix A lists SourceFinder’s 59 filters.)

- **User Interface** — This component allows test developers to browse the Candidate...
Source Database, to search for documents with specific feature values, and to record comments for viewing by other test developers.

Figure 3. A high-level overview of the SourceFinder system (taken from Bauer & Jha, 1999; reprinted with permission of the author).

SourceFinder System Processing

Most of SourceFinder’s innovative processing takes place in the Text Gatherer and the Text Selector. While the Text Gatherer was designed to function as a Web crawler, it can also accommodate paper-based source materials if these documents are scanned into the Text Database. The Text Gatherer can then be instructed to access that specific database, allowing for inexpensive screening of large quantities of paper-based source material. However, the advantages of accessing source materials that exist only in paper may be offset by the expenses associated with operating a scanner and insuring the quality of the scanned text.

The text features implemented in the SourceFinder prototype were intended to allow users to screen source text for length, content, clarity, density, and level of argumentation. Length filtering is extremely important, since copyright constraints dictate that the number of words in a reading comprehension passage must be less than 10% of the total number of words in the source document from which the passage was extracted. For example, the source document for a 450-word GRE passage must include at least 4,500 words. The length filter in the
existing SourceFinder prototype operates by counting all of the words in the retrieved document — including titles, footnotes, and references. SourceFinder accepts only those documents that are long enough to provide a passage of the required length (250 words or 450 words) without violating the copyright constraint. Documents that are either too short or too long are rejected.

SourceFinder’s content filter assigns each retrieved document to one of five different content categories: humanities, social sciences, physical sciences, biological sciences or “none of the above.” To classify a document, SourceFinder compares the words found in the document to a previously developed list of content indicator words.

SourceFinder screens for clarity, density, and level of argumentation using a predefined set of text features. SourceFinder’s feature set includes both superficial text features (including average sentence length, number of unique words per sentence, number of sentences per paragraph, and percentage of determiners — words like “the,” “a,” “some,” “these”), and more abstract features that depend on knowledge not available within the document itself (such as word frequency indices — average word difficulty, estimated readability level, and percentage of key words; Carroll, Davies & Richman, 1971; Chall & Dale, 1997). SourceFinder also uses an independently defined key word list to locate texts that exhibit conceptual conflicts. The key word list includes words or phrases that have been found to occur frequently in such texts — such as “however,” “arguably,” “in contrast,” “competing,” “disputable,” and “incompatible.” (Appendix B provides all of the words in SourceFinder’s key word list.)

Document screening is accomplished by comparing the feature values calculated for each new document to a predetermined set of feature acceptability thresholds. A document is classified as “acceptable” only if all of its feature values fall within the specified thresholds. A single feature value that falls outside the relevant acceptability threshold will cause a document to be rejected.

The feature acceptability thresholds used in the original prototype were determined from a database of 116 disclosed GRE reading passages. Because a database of “rejected” sources was not available at the time that the acceptability thresholds were determined, it is possible that these thresholds may not be optimally configured to discriminate between acceptable and unacceptable sources.
**1999 Preliminary Evaluation**

A limited evaluation of the SourceFinder prototype was conducted in 1999 by Bauer and Jha. Six documents that met all of SourceFinder’s screening requirements were extracted from a collection of history articles found on the Internet. Two expert test developers were asked to rate each document. The test developers agreed that three of the six documents were appropriate for use as source documents for GRE passages in the social sciences content area. Because this evaluation included only six documents, the relationship between source acceptability and calculated feature values was not investigated. The authors of the study noted that “[w]hile this brief trial evaluation is too short to make strong claims, it does show the potential of the SourceFinder” (Bauer & Jha, 1999, p. 9).

**Method**

The objectives of the current study were accomplished by running a sample of electronic documents in each of the GRE-specified content areas through SourceFinder and comparing the judgments of expert GRE test developers to SourceFinder’s source acceptability classifications. Modifications were made to the SourceFinder prototype to facilitate this study. Also, a corpus was created and documents selected from that corpus were evaluated by the test developers. Feedback from the raters informed changes to SourceFinder’s search capability and interface design, as well as other system improvements.

**Modifications to the SourceFinder Prototype**

Prior to this study, SourceFinder was implemented as a stand-alone system on a desktop computer, and the sources it collected were held locally on that machine. From an operational perspective, it was recognized that a shift to client-server architecture would be more efficient and would enable greater communication between item writers. For example, this modification would allow item writers to access SourceFinder’s database of candidate passages from their individual work stations, while also allowing a “source manager” to use SourceFinder to create large, quality collections of sources held in a centralized passage database on the server machine. (For the GRE program, one item writer plays the role of source manager for paper-based sources.) Shifting to a centralized server would also enable SourceFinder to continuously search the Web for source material without drawing computational resources from individual item writer’s computers. Thus, the system modifications needed to run SourceFinder in a client-server
mode were implemented as part of this project.

When the test developers explained that their experience has led them to prefer certain journals as text sources, the researches attempted to deploy SourceFinder as a database explorer. Although this capability was not programmed into the SourceFinder prototype, a database explorer implementation was simulated by transferring a large number of candidate source documents from the EBSCO Information Services Web site to a database, and then having SourceFinder process the entire pool. This effort permitted the research staff to examine the feature values of documents that SourceFinder rejects as well as those it accepts.

Finally, an experimental version of the SourceFinder prototype was developed for use in selecting individual documents to be included in the corpus. This experimental version differed from the original prototype in that individual filters could be turned on or off at will. This enhanced filtering capability enabled research staff to restrict the corpus to only those documents that (a) were classified by SourceFinder as belonging to one or another of the four main GRE content areas and (b) met or exceeded the length constraint.

**Corpus Creation**

A moderately sized corpus of candidate source documents was constructed for use in this and future SourceFinder evaluations. The set of journals that had been deemed “capable of yielding usable GRE source material” by GRE test developers provided the database of 1,054 source documents considered during corpus creation. (Appendix C lists these journals.) These electronic articles were aggregated into a corpus, as follows:

*Step 1. Selection with content filtering only.* In this initial step, SourceFinder’s operational content filter was used to classify each retrieved document into one of the five nonoverlapping content areas mentioned earlier — humanities, social sciences, physical sciences, biological sciences, or none of the above. Documents classified as “none of the above” were labeled “unusable” and were removed from further consideration.

*Step 2. Further selection using the length filter.* Next, the documents retrieved in step 1 were filtered for length. Documents that failed to meet the length constraint were labeled “unusable” and were removed from further consideration.
Step 3. Evaluation using the clarity, density, and argumentation feature set. To insure that the completed corpus would include both acceptable and unacceptable source documents, in this step all documents that met the length constraint were marked for inclusion in the corpus. Each of these documents was then evaluated using SourceFinder’s clarity, density, and argumentation feature sets.

Step 4: Final source acceptability classifications. In this last step, documents were classified as either acceptable or unacceptable.

To provide a more detailed accounting of SourceFinder’s filtering capabilities, two separate SourceFinder acceptability classifications were developed for each document. The first classification was implemented with a relatively loose set of filtering thresholds, and the second was implemented with a slightly tighter set of threshold values. The latter set of threshold values was developed by tightening three specific threshold values while leaving all others unchanged. Table 1 lists the two sets of threshold values used to implement loose and tight filtering.

Table 1
Two Sets of Threshold Values for SourceFinder

<table>
<thead>
<tr>
<th>Filter</th>
<th>Threshold value</th>
<th>Loose</th>
<th>Tight</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniqueWordperWord — the number of unique words in the document divided by the total number of words in the document</td>
<td>.12</td>
<td>.16</td>
<td></td>
</tr>
<tr>
<td>AvgParaLength — the total number of words in the document divided by the total number of paragraphs in the document</td>
<td>250</td>
<td>280</td>
<td></td>
</tr>
<tr>
<td>AvgParaDifficulty — a text readability measure averaged over all paragraphs</td>
<td>50</td>
<td>90</td>
<td></td>
</tr>
</tbody>
</table>

The filters were implemented in a noncompensatory manner. That is, a document was classified as acceptable only if all of its feature values fell above the relevant thresholds. A single feature value below the relevant threshold triggered an unacceptable classification. This four-step approach yielded a 142-document corpus with an approximately equal distribution in each of the four main GRE content categories and complete feature information for each
document. These data can be used in future research to compare SourceFinder performance to human performance and to address the possibility that documents rejected by SourceFinder might be accepted by humans. The corpus also provides information about the distribution of acceptable and unacceptable documents at relevant URLs on the Web.

**Document Selection**

Due to human resource constraints, not all of the 142 documents in the corpus could be included in the expert evaluation portion of this study. Consequently, a design for sampling documents from the corpus was developed by considering the following information:

1. *Expected false-positive rate.* Based on the preliminary evaluation reported in Bauer and Jha (1999), it was expected that SourceFinder would have a high false-positive rate — that is, that many of the documents that SourceFinder classified as acceptable would subsequently be rated as unacceptable by the expert test developers.

2. *Expected false-negative rate.* Although SourceFinder’s false-negative rate (i.e., the probability that SourceFinder would reject a document that an expert test developer would accept) was expected to be low, no concrete data about this rate had ever been collected. Consequently, a small pilot study was designed and implemented to gather information about SourceFinder’s false-negative rate. First, a sample of 20 documents that SourceFinder had classified as unacceptable under the loose constraints was selected. All 20 documents had met the length constraints and had been classified into a content category, but had failed to meet one or more of SourceFinder’s other constraints. Thus, SourceFinder had classified all 20 documents as being unacceptable for reasons other than length and content. Next, a single test developer was asked to evaluate each document. Two of the 20 documents were judged as being acceptable for use as a GRE source document. Based on these results, a relatively low false-negative rate was expected.

3. *Differences in the losses associated with false-negative and false-positive decisions.* SourceFinder’s primary mission is to help test developers locate more usable sources in less time. Since no test development time is lost if SourceFinder passes over one acceptable document on its way to locating a different acceptable document, there is very little loss associated with a false-negative decision. False-positive decisions, on the other hand, incur serious losses; each time SourceFinder presents an unacceptable document to a test developer for review, valuable test developer time is lost.

4. *Time required for expert evaluation of a typical source document.* Information about the time needed for expert evaluation of a typical source document was collected in the 20-document pilot study described above. These results suggested that, for approximately two thirds of the documents, about six minutes per document were needed to complete the evaluation. For the remaining one third, however, as much as 30 minutes per document was needed.
5. **Interrater agreement:** Based on previous experience with judged evaluation data, the source suitability judgments provided by different experts were expected to exhibit a moderate amount of disagreement.

Based on the information above, it was determined that multiple human ratings would be needed for each document and that resource constraints would allow for a total of 240 document evaluations. Consequently, the final design called for (a) sampling 20 documents in each of the four content areas, (b) selecting all 80 documents from those that SourceFinder had classified as acceptable under the loose constraints, and (c) collecting three human ratings for each document. This strategy facilitated subsequent analyses of SourceFinder performance under both loose and tight constraints.

**Allocation of Documents to Raters**

Since a total of four expert test developers agreed to participate in the study, the 20 documents in each content area were grouped into sets that each included five documents. Each rater received a total of 15 documents (three sets) in each of the four content areas. The design for allocating documents to raters in the humanities content area is shown in Table 2. The same design was used in each of the three remaining content areas. This design insured that each rater evaluated a total of 60 documents (15 in each content area) and that each pair of raters evaluated 40 documents in common.

**Table 2**

*Design for Allocating Documents to Raters In the Humanities Content Area*

<table>
<thead>
<tr>
<th>Document set</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Total raters per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (1-5)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>3</td>
</tr>
<tr>
<td>B (6-10)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>3</td>
</tr>
<tr>
<td>C (11-15)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>3</td>
</tr>
<tr>
<td>D (16-20)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total documents rated</strong></td>
<td><strong>15</strong></td>
<td><strong>15</strong></td>
<td><strong>15</strong></td>
<td><strong>15</strong></td>
<td></td>
</tr>
</tbody>
</table>
Rater Instructions

Interviews with the test developers prior to the start of the evaluation indicated that some source documents can be extremely difficult to evaluate. For such “borderline” documents, an accurate assessment of source suitability almost requires taking the time to craft a passage and author the required numbers of items. If all of the required items can be written, the source is deemed acceptable. When one or more of the required items cannot be written, the source is considered unacceptable. For this reason, a set of instructions based on a “triage” method was designed and piloted with one test developer and refined for use in this study.

The triage method was designed to maximize the quality of test developer judgments while minimizing demands on test developer time. The method called for sorting “easy” documents first, in order to insure that all judgments that could be made quickly would be made first. Understanding that some documents might be less clearly acceptable than others, the system employed a “probable accept” category. A “probable reject” category was also included to capture the occasional situation in which a second or third look was taken at a document that would require extensive revision to create a passage, but had not yet been rejected simply because more amenable sources may not be found.

Rater instructions specified that the test developers should spend a relatively short amount of time identifying documents that were easy to evaluate, and then return to more problematic documents later. During the first review, raters were to classify each document as “definite accept,” “definite reject,” or “don’t know yet.” During the second review, documents designated “don’t know yet” in the first round were to be categorized as “definite accept,” “probable accept,” “definite reject,” “probable reject,” or “undecided.” This classification strategy is illustrated in Figure 4. In addition test developers were asked to provide brief, open-ended comments explaining their reasons for assigning particular documents to particular categories. The resulting set of instructions used in the full evaluation is shown in Appendix D.
Results

Corpus Creation

Table 3 summarizes the results of the corpus creation effort. Each row of the table lists the number of documents retained as acceptable in each content area after a particular SourceFinder filter was applied. The first row shows that, after content filtering, the online database of 1,054 documents yielded approximately 200 documents in each of the four main content areas, or a total of 840 documents. The second row confirms that many of these documents had to be rejected simply due to inappropriate length. The effects of the clarity, density and argumentation filters are shown in the third row.

As can be seen, only 24% of the humanities documents that met the content and length requirement were classified as having appropriate levels of clarity, density, and argumentation. In the other three content areas, the percentage of documents classified as having appropriate levels of clarity, density, and argumentation was somewhat higher, averaging about 44%. Based on the test developer’ comments, these three filters are the keys to finding appropriate sources.
Table 3

Summary of Classification Results Obtained During Corpus Creation

<table>
<thead>
<tr>
<th>Classification and type of filtering</th>
<th>Humanities</th>
<th>Social sciences</th>
<th>Physical sciences</th>
<th>Biological sciences</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptable after content filtering only</td>
<td>247</td>
<td>181</td>
<td>176</td>
<td>236</td>
<td>840</td>
</tr>
<tr>
<td>Acceptable after content and length filtering</td>
<td>153</td>
<td>101</td>
<td>49</td>
<td>89</td>
<td>392</td>
</tr>
<tr>
<td>Acceptable after content, length, clarity, density, and argumentation filtering</td>
<td>36</td>
<td>45</td>
<td>23</td>
<td>38</td>
<td>142</td>
</tr>
<tr>
<td>Pass rate for clarity, density, and argumentation filters</td>
<td>24%</td>
<td>45%</td>
<td>47%</td>
<td>43%</td>
<td>36%</td>
</tr>
</tbody>
</table>

Note. In rows one through three, results are provided in numbers of documents. The percentages in row four represent the proportion of those documents meeting content and length requirements that subsequently also passed for clarity, density, and argumentation.

Evaluation of SourceFinder Against Human Judgments

Analysis of Expert Acceptability Judgments

Expert acceptability judgments were collected for a total of 100 documents — 20 documents that SourceFinder rejected using loose constraints and 80 documents that SourceFinder classified as acceptable using loose constraints. By design, all 100 documents were selected from those that had met SourceFinder’s length constraints. Thus, all rejected documents were classified as unacceptable for reasons other than length.

Since it was expected, and verified, that SourceFinder’s false-negative rate would be quite low, and since guarding against false-negative decisions was not a primary goal of the study, a single rater evaluated the subset of documents that SourceFinder rejected. The remaining documents (i.e., those SourceFinder found acceptable) were all rated by three different raters.

In some cases, the classification output received from the test developers did not employ the category labels described in Figure 4 and in the rater instructions. Instead, the test developers provided free-form comments that had to be translated, after the fact, into one of the five acceptability categories. Table 4 provides a sample of these comments and the subsequent classifications assigned to them.
Table 4

Sample Free-Form Comments and Subsequent Classifications

<table>
<thead>
<tr>
<th>Comment</th>
<th>Subsequent classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) “Nice but may require pretty extensive cobbling.”</td>
<td>Probable accept</td>
</tr>
<tr>
<td>(b) “Just don’t see an argument that is of the right scale for a short, which is all that an article of this length could sustain.”</td>
<td>Probable reject</td>
</tr>
<tr>
<td>(c) “Not promising but perhaps not impossible.”</td>
<td>Undecided</td>
</tr>
<tr>
<td>(d) “Fairly technical, and content is overly generalized in some sections, and overly specific in others.”</td>
<td>Probable reject</td>
</tr>
<tr>
<td>(e) “This might work for a short.”</td>
<td>Probable accept</td>
</tr>
</tbody>
</table>

The authors of this report translated free-form comments to specific acceptability categories. Although this solution was not ideal, no other response was feasible given the time constraints within which the study had to be completed. However, as indicated by the above sample of comments, translations were only needed for documents that eventually wound up being classified into one of the three innermost acceptability categories. That is, all documents classified as “definite accept” or “definite reject” were clearly labeled as such by the human raters.

Table 5 summarizes the resulting source acceptability judgments, by rater. The four right-most columns of the table characterize the experts’ evaluations of the 80 documents that SourceFinder had classified as acceptable. An acceptance rate was calculated for each rater by collapsing over the “definite accept” and the “probable accept” categories. As can be seen, the acceptance rates provided by the experts ranged from a low of 7% to a high of 30%. In interpreting these results, it is important to recall that some judgments are translations of free-form comments and also that different raters evaluated different subsets of documents. Thus, the observed differences include rater-to-rater variation, classification error, and some amount of document-to-document variation.
Table 5
Source Acceptability Ratings, by Rater, for 100 Potential Source Documents

<table>
<thead>
<tr>
<th>Rater decision</th>
<th>20 SourceFinder rejects</th>
<th>80 documents rated by SourceFinder as acceptable using loose constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rater 1</td>
<td>Rater 1</td>
</tr>
<tr>
<td>Definite reject</td>
<td>13</td>
<td>37</td>
</tr>
<tr>
<td>Probable reject</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Undecided</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Probable accept</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Definite accept</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Total documents rated</td>
<td>20</td>
<td>60</td>
</tr>
<tr>
<td>Acceptance rate&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.10</td>
<td>.13</td>
</tr>
</tbody>
</table>

<sup>a</sup> Includes both “definite accept” and “probable accept.”

Table 6 summarizes the acceptability judgments provided by each pair of raters. It lists the percentage of documents that were jointly rated by each pair of raters as either unacceptable (i.e., “definite reject” or “probable reject”) or acceptable (i.e., “definite accept” or “probable accept”). For example, the table shows that raters 1 and 2 jointly classified a total of 26 documents as unacceptable. Overall, the table shows that all pairs of raters agreed that this set of 80 SourceFinder “accepts” contained a high proportion of false positives. In particular, the false-positive rate ranged from a low of 55% for raters 3 and 4 to a high of 65% for raters 1 and 2.

In interpreting this level of variation it is important to recall that no document was rated by all four raters. Thus, the subset of 40 documents that was jointly rated by raters 1 and 2 did not include any documents that were also included in the subset of 40 documents that was jointly rated by raters 3 and 4. Given this fact, it is clear that the current results are fairly consistent and provide useful information about the percentage of false positives to be expected in any future sample of documents accepted by SourceFinder. In particular, in any future sample of 40 SourceFinder “accepts,” we should expect to find about 23 documents that would be jointly classified as unacceptable by any pair of raters. This is a joint false-positive rate of about 57.5%.
Table 6

**Raters’ Joint Classification Results for 80 SourceFinder “Accepts”**

<table>
<thead>
<tr>
<th>Pairs of raters</th>
<th>Number of documents rated in common</th>
<th>Documents jointly classified as either “definite reject” or “probable reject”</th>
<th>Documents jointly classified as either “definite accept” or “probable accept”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Number</td>
<td>Percent</td>
</tr>
<tr>
<td>Raters 1 and 2</td>
<td>40</td>
<td>26</td>
<td>65.0</td>
</tr>
<tr>
<td>Raters 1 and 3</td>
<td>40</td>
<td>23</td>
<td>57.5</td>
</tr>
<tr>
<td>Raters 1 and 4</td>
<td>40</td>
<td>23</td>
<td>57.5</td>
</tr>
<tr>
<td>Raters 2 and 3</td>
<td>40</td>
<td>23</td>
<td>57.5</td>
</tr>
<tr>
<td>Raters 2 and 4</td>
<td>40</td>
<td>23</td>
<td>57.5</td>
</tr>
<tr>
<td>Raters 3 and 4</td>
<td>40</td>
<td>22</td>
<td>55.0</td>
</tr>
</tbody>
</table>

Table 6 also provides information about the number and proportion of true positives found by each pair of raters. As can be seen, very few true positives were found. The table also suggests that the proportion of true positives was higher in some samples than in others. This result should not be surprising given the scarcity of true positives detected overall. In particular, once we recall that the subset of documents evaluated by raters 2 and 3 did not include any of the documents included in the subset reviewed by raters 1 and 4, we should not be surprised to learn that raters 2 and 3 found only one true accept, while raters 1 and 4 found six. What is surprising, however, is that four times out of six, a subset of 40 documents rated by SourceFinder as acceptable included three or fewer truly acceptable documents, and that six times out of six, a subset of 40 documents rated by SourceFinder as acceptable included fewer than eight truly acceptable documents. These results suggest that in any future sample of 40 documents accepted by SourceFinder, expert test developers are not likely to find a high percentage of acceptable documents.

Additional information about the judged acceptability status of each document in the sample of 80 SourceFinder “accepts” was obtained by combining the three ratings provided for each document, as follows:
When all three raters reported the same decision for a document, that decision was taken as the combined judgment. For example, if all three categorized a document as a “definite reject (DR),” the decision was as follows: (DR, DR, DR) → DR.

When two of the three raters reported the same decision for a document, that decision was also taken as the combined judgment. That is, the third rating, no matter what it was, was dropped. For example, a document rated a “definite reject (DR)” by two raters and “undecided (UN)” by a third was judged as follows: (DR, DR, UN) → DR.

When all three raters reported different decisions for a document, the decisions were ordered as shown in the decision process graphic in Figure 4, and the middle decision was taken as the combined judgment. For example, a document rated “definite reject (DR),” “undecided (UN),” and “definite accept (DA)” by three raters would be resolved as follows: (DR, UN, DA) → UN.

The resulting judgments were termed “consensus judgments.”

Table 7 summarizes all consensus judgments by content area. For purposes of comparison, the table also lists the judgments collected for the subset of 20 documents that had been rejected by SourceFinder. The table shows that, in the sample of 80 SourceFinder “accepts,” a total of 45 documents (56%) were judged as definitely unacceptable by two or more expert test developers, and an additional 13 documents (16%) were judged as probably unacceptable by two or more expert test developers. Thus, the acceptance rates obtained in the current evaluation are lower than those reported in the Bauer and Jha (1999) evaluation. In addition, the data suggest that SourceFinder is somewhat more successful at locating acceptable text in the biological and physical sciences than in the humanities and social sciences content areas. This maybe due to the impact of vocabulary differences, the likelihood of first person narratives, the lack of a standard format for presenting arguments, or other reasons.
Table 7

**Consensus Judgments of Four Expert Human Raters by Content Area**

<table>
<thead>
<tr>
<th>Expert judgment&lt;sup&gt;a&lt;/sup&gt;</th>
<th>SourceFinder rejects</th>
<th>Articles rated as acceptable by SourceFinder using loose constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mixed</td>
<td>Humanities</td>
</tr>
<tr>
<td>Definite reject</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Probable reject</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Undecided</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Probable accept</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Definite accept</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total documents rated</strong></td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td><strong>Acceptance rate&lt;sup&gt;a&lt;/sup&gt;</strong></td>
<td>.10</td>
<td>.10</td>
</tr>
</tbody>
</table>

<sup>a</sup> Includes both “definite accept” and “probable accept.”

It is useful to consider the extent to which judgments provided by individual raters agreed with the calculated consensus judgments. Table 8 provides the percentage of times that individual raters agreed with consensus judgments in each content area — either exactly or with no more than a one-point difference. The table also provides both exact and adjacent agreement rates adjusted for expected agreement due to chance. Adjusted agreement rates were calculated using Cohen’s kappa (Cohen, 1960). The table confirms that, even after adjusting for chance agreement, on average 87% of the individual judgments were within one point of the reported consensus judgments. The data also suggest that rater agreement in the humanities and social sciences content areas was slightly higher than that in the biological and physical sciences content areas.

The extent to which each individual rater agreed with the calculated consensus judgments was also considered. As shown in Table 9, raters 1 and 2 were somewhat more likely to agree with calculated consensus judgments.
Table 8

Individual Agreement With Consensus Judgments of Source Acceptability

<table>
<thead>
<tr>
<th></th>
<th>Percent exact agreement</th>
<th>Percent exact or adjacent agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not adjusted</td>
<td>Kappa adjusted for random agreement</td>
</tr>
<tr>
<td>Humanities</td>
<td>75</td>
<td>69</td>
</tr>
<tr>
<td>Social sciences</td>
<td>73</td>
<td>67</td>
</tr>
<tr>
<td>Biological sciences</td>
<td>67</td>
<td>58</td>
</tr>
<tr>
<td>Physical sciences</td>
<td>73</td>
<td>67</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>65</td>
</tr>
</tbody>
</table>

Table 9

Individual Agreement With Consensus Judgments of Source Acceptability by Rater

<table>
<thead>
<tr>
<th></th>
<th>Percent exact agreement</th>
<th>Percent exact or adjacent agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not adjusted</td>
<td>Kappa adjusted for random agreement</td>
</tr>
<tr>
<td>Rater 1</td>
<td>77</td>
<td>71</td>
</tr>
<tr>
<td>Rater 2</td>
<td>78</td>
<td>73</td>
</tr>
<tr>
<td>Rater 3</td>
<td>63</td>
<td>54</td>
</tr>
<tr>
<td>Rater 4</td>
<td>70</td>
<td>63</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>65</td>
</tr>
</tbody>
</table>

Precision and Recall of SourceFinder’s Classification Algorithm

In many information retrieval applications, system performance is characterized using two particular measures: precision and recall (Litman & Passonneau, 1995; van Rijsbergen, 1979). Precision is defined as the proportion of acceptable classifications generated by the system that were later judged to be correct classifications by human experts. For the current application, precision is calculated by dividing the number of documents that SourceFinder correctly classified as acceptable (i.e., true accepts only) by the total number of documents that it
classified as acceptable (i.e., the sum of true accepts and false accepts). Recall is defined as the proportion of true Accepts that the system successfully located in the corpus. For the current application, recall is calculated by dividing the number of documents that SourceFinder correctly classified as acceptable by the total number of truly acceptable documents present in the corpus.

The data in Table 7 can be used to estimate the precision and recall of SourceFinder’s source classification algorithm when the loose threshold values are in effect. As Table 10 illustrates, after collapsing over the four content areas, 14 of the 80 documents SourceFinder rated as acceptable in Table 5 were true positives. That is, 14 of these 80 documents were also classified as acceptable or probably acceptable by human experts. Thus, the proportion of true accepts out of all SourceFinder’s loose accepts (precision) is 14/80 or 17.5%. Stated another way, when SourceFinder’s loose thresholds are in effect, 17.5% of the documents that SourceFinder classifies as acceptable are likely to also be acceptable to human test developers.

Setting the constraints at this level also resulted in a high false-positive rate of 82% (66/80 = .825). In other words, 82% of the time, a document that SourceFinder classified as acceptable was later found to be unacceptable by test developers. SourceFinder’s ratio of “bad” (false positives) to “good” (true positives) is 66/14, or 4.7, meaning that on average, a test developer will likely have to read through about five unacceptable sources to locate each new acceptable source. This result is similar to the current manual process where an average of 10 sources might yield two usable sources for passage creation.

**Table 10**

*Classification Results With Loose Acceptance Criteria*

<table>
<thead>
<tr>
<th>SourceFinder accepted</th>
<th>Human raters accepted</th>
<th>Human raters rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>14</td>
<td>66</td>
</tr>
<tr>
<td>SourceFinder rejected</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>84</td>
</tr>
</tbody>
</table>
The precision of SourceFinder’s tight threshold set was also calculated. As Table 11 indicates, when SourceFinder’s tight thresholds are employed, the total number of documents rated as acceptable by SourceFinder drops to 45, and of these, a total of six are true positives. The precision of SourceFinder’s tighter acceptance algorithm is 6/45, or about 13%. That is, when SourceFinder’s tight thresholds are in effect, 13% of the documents that SourceFinder classifies as acceptable are likely to also be acceptable to human test developers.

Setting the constraints at this level also resulted in a slightly higher false-positive rate of about 87% (39/45 = .866). In other words, 87% of the time, a document that SourceFinder classified as acceptable was later found to be unacceptable by test developers. SourceFinder’s ratio of “bad” (false positives) to “good” (true positives) is 39/6, or 6.5, meaning that on average, a test developer will likely have to read through about six unacceptable sources to locate each new acceptable source. Thus, the data suggest that a strategy of tightening the existing SourceFinder screening criteria is not likely to improve the precision of SourceFinder’s source classification algorithm.

Table 11
Classification Results With Tight Acceptance Criteria

<table>
<thead>
<tr>
<th>Human raters accepted</th>
<th>Human raters rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>SourceFinder accepted</td>
<td>6</td>
</tr>
<tr>
<td>SourceFinder rejected</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>16</td>
</tr>
</tbody>
</table>

Recall values were also estimated for this application. As Table 10 shows, human experts rated 16 of the 100 documents they evaluated as true positives. That is, a total of 16 documents were classified by test developers as being either acceptable or probably acceptable. When SourceFinder was implemented with loose acceptance criteria, it successfully located 14 of these 16 true positives. Thus, under loose threshold values, SourceFinder’s recall measure is 14/16 or 87.5%. When SourceFinder was implemented with tight acceptance criteria, the number of true positives it located dropped to six. Thus, when SourceFinder’s tight acceptance criteria are in
effect, recall is 6/16 or about 38%. As was the case for the precision estimates summarized above, these results suggest that a strategy of tightening SourceFinder’s existing screening criteria is not likely to result in improved performance relative to human judgments.6

**Rater Comments**

For each of the 80 documents SourceFinder rated as acceptable, individual raters provided comments on the aspects of source variation that contributed to their acceptability judgments. As a first step in analyzing these data, a comment classification system was developed. As Table 12 shows, all of the comments test developers offered for the 80 documents SourceFinder rated as acceptable were classified into one or more of seven categories.

Table 13 presents sample rater comments for each of these seven categories. These comments illustrate the individual words and phrases expert test developers use when evaluating candidate source documents. As you will note, certain types of comments tended to co-occur. Sample source documents illustrating each type of comment are provided in Appendix E.

**A Simulation of Serial Filtering**

Since rejection features tended to co-occur, success at detecting some undesirable document features may also reduce the incidence of other nontargeted features. Table 14 illustrates the extent to which this might be the case. The table simulates a serial filtering process. Each cell shows the number of documents that were classified as unacceptable due to a specified aspect of the text; documents classified at earlier stages were not reclassified.

The first column of the table shows that in the subsample of 20 humanities documents SourceFinder classified as acceptable, expert test developers found a total of eight that were the wrong genre, an additional five that were judged to be in violation of the accessibility constraint, an additional three that were classified as exhibiting problematic reasoning, and an additional two that were judged to be in violation of sensitivity guidelines. Interestingly, no document was judged to be inappropriate due to tone/style or topicality/datedness *only*. Similar results were obtained in the other three content areas. These results suggest that additional system development work is most needed in the areas of genre detection, accessibility evaluation, and reasoning/argumentation evaluation.
Table 12

_A Classification System for Source Variation_

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Percent of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problematic reasoning</td>
<td>The source contains argumentative reasoning, but it might be difficult to translate into a passage as a result of one or more of the following problems: too many interconnected argument threads to extract a single clear argument; argument is too dispersed throughout the text to be worth the effort; argument lacks clarity, density, or substance.</td>
<td>69%</td>
</tr>
<tr>
<td>Wrong genre</td>
<td>The source does not provide a scholarly analysis. Includes documents that are primarily descriptive or expository, as well as first-person narratives, memoirs, interviews, biographies, and guidebooks.</td>
<td>23%</td>
</tr>
<tr>
<td>Inaccessible</td>
<td>The source is too specialized, too full of jargon, too technical, too abstract, or too dependent on a close reading of another work, such as a novel, poem, or film. Would not be accessible to the entire GRE population.</td>
<td>28%</td>
</tr>
<tr>
<td>Inappropriate tone or style</td>
<td>Tone is too casual or too flip, or author is too strongly opinionated.</td>
<td>7%</td>
</tr>
<tr>
<td>Topicality or datedness issues</td>
<td>Contains material that would quickly become dated.</td>
<td>3%</td>
</tr>
<tr>
<td>Sensitivity issues</td>
<td>Contains sensitive or inflammatory content.</td>
<td>2%</td>
</tr>
<tr>
<td>On target</td>
<td>The source provides the desired mix of analysis/explanation at the desired tone with the desired level of accessibility and with the desired frequency (i.e., length).</td>
<td>9%</td>
</tr>
</tbody>
</table>

*Note.* Because a given document could be rejected for more than one reason, the percentages shown in the table do not necessarily sum to 100%.
### Table 13

**Sample Rater Comments by Category**

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problematic reasoning</td>
<td>“Would require extensive cobbling,” “not easily summarized,” “hard to condense,” “a distillation of the arguments might be problematic”</td>
</tr>
<tr>
<td>Wrong genre</td>
<td>“First-person, informal narrative, descriptive rather than argumentative,” “brief descriptive summaries, no argument develops,” “descriptive summary, no sustained argument,” “Content is descriptive in nature, don’t see useful theories/arguments/explanations, etc.,” “brief description of a museum show, … no real analysis/theory, etc.”</td>
</tr>
<tr>
<td>Inaccessible</td>
<td>“Too specialized,” “too abstract and jargon filled,” “might require too much glossing of terminology,” “There is an argument here, but it is way too specialized with lots of theoretical concepts and buzzwords, impossible to summarize easily,” “too esoteric for most examinees,” “think this article requires a level of cyber-sophistication that many examinees might not possess,” “a lot of technical language and much abstraction”</td>
</tr>
<tr>
<td>Inappropriate tone or style</td>
<td>“Style is breezier and more casual than GMAT/GRE generally use, but the article does contain some arguments,” “Contains some manipulable concepts, but tone is too flip — would have to be thoroughly recast,” “full of stereotypical generalities,” “strongly opinionated”</td>
</tr>
<tr>
<td>Topicality or datedness issues</td>
<td>“Highly topical,” “would become dated”</td>
</tr>
<tr>
<td>Sensitivity issues</td>
<td>“Contains clearly inflammatory material — vegetarianism and question of animal rights,” “emotionally charged subject (hate crimes),” “Examples range from Playboy nudes to communion as analogous to cannibalism”</td>
</tr>
<tr>
<td>On target</td>
<td>“Expresses a definite viewpoint in opposition to a conventional/wide-spread viewpoint,” “lots of good tension and many possibilities to exploit,” “appears to have plenty of arguments/theories/explanations and counter-arguments/theories/explanations, all of which appear to be accessible”</td>
</tr>
</tbody>
</table>
Table 14

A Simulation of Serial Filtering by Content Area

<table>
<thead>
<tr>
<th></th>
<th>Humanities</th>
<th>Social sciences</th>
<th>Biological sciences</th>
<th>Physical sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total documents to be classified</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Classification results:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Wrong genre</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>2. Inaccessible</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>3: Problematic reasoning</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4. Inappropriate tone or style</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5. Sensitivity issues</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>6. Topicality or datedness issues</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total documents classified</td>
<td>18</td>
<td>18</td>
<td>16</td>
<td>14</td>
</tr>
</tbody>
</table>

Note. Documents classified at earlier stages were NOT reclassified at later stages.

Evaluation of Individual SourceFinder Features

The preceding evaluation of SourceFinder’s decisions against those of human raters indicated that SourceFinder’s success at locating documents that are acceptable to expert test developers requires additional research and development in the areas of genre detection, accessibility evaluation, and reasoning classification. Modifications to existing SourceFinder features could provide a basis for improvement in these areas. Analyses of current SourceFinder features and potential modifications are discussed below.

First-Person Pronouns as a Genre Filter Enhancement

Further analysis of the raters’ comments indicated that five documents in the social science content area were rejected because they were first-person narratives. Thus, a tendency towards first-person narration accounts for roughly 28% of the total number of social sciences documents that were rejected. This observation provides an illustration of how more careful analysis of individual features could enhance SourceFinder’s performance: The additional filtering capability needed to robustly exclude documents with an unusually high proportion of first-person pronouns (e.g. “I”, “my”, “mine”) is relatively easy to implement.

One of SourceFinder’s existing features, PronounPerWord, looks at the proportion of all
types of pronouns in a document. This feature is too coarse-grained to help identify first-person narratives. However, as an initial determination of the feasibility of automatically detecting first-person narratives, simple ratios of counts of the single pronoun “I” to the total number of words were computed for 14 sample documents.7

Each of the 14 sample documents was assigned to one of three test sets:

- Test set A consisted of five documents that human raters classified as both unacceptable and narrative.
- Test set B consisted of four documents that human raters classified as both unacceptable and nonnarrative.
- Test set C included five documents that were classified as both acceptable and nonnarrative.

Table 15 shows the resulting density of the first-person pronoun “I” in each sample document. The table suggests that documents classified as narrative tend to include a higher density of the first-person pronoun “I” than documents classified as nonnarrative. In particular, the set of five narrative documents considered here contained an average of 8.6 occurrences of this pronoun per 1,000 words. In contrast, rejected documents classified as nonnarrative used “I” an average of only 1.7 times per 1000 words, and accepted documents classified as nonnarrative contained an average of 1.1 instances of the first-person pronoun “I” per 1000 words. This suggests that a filter designed to detect high proportions of first-person pronouns may help to reduce SourceFinder’s false-positive rate. Also, within each category, humanities documents differed most from the mean. This result suggests that different threshold values may be needed for humanities documents.
Table 15

Potential of First-Person Pronoun “I” as an Additional SourceFinder Filter

<table>
<thead>
<tr>
<th>Document</th>
<th>Content area</th>
<th>Raters’ acceptability classification</th>
<th>Raters’ narrative classification</th>
<th>Incidence of first-person pronoun “I” per 1000 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-1</td>
<td>HU</td>
<td>Reject</td>
<td>Narrative</td>
<td>13.3</td>
</tr>
<tr>
<td>A-2</td>
<td>HU</td>
<td>Reject</td>
<td>Narrative</td>
<td>13.1</td>
</tr>
<tr>
<td>A-3</td>
<td>SS</td>
<td>Reject</td>
<td>Narrative</td>
<td>6.7</td>
</tr>
<tr>
<td>A-4</td>
<td>SS</td>
<td>Reject</td>
<td>Narrative</td>
<td>4.1</td>
</tr>
<tr>
<td>A-5</td>
<td>BS</td>
<td>Reject</td>
<td>Narrative</td>
<td>5.8</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td></td>
<td>8.6</td>
</tr>
<tr>
<td>B-1</td>
<td>HU</td>
<td>Reject</td>
<td>Nonnarrative</td>
<td>4.1</td>
</tr>
<tr>
<td>B-2</td>
<td>SS</td>
<td>Reject</td>
<td>Nonnarrative</td>
<td>1.0</td>
</tr>
<tr>
<td>B-3</td>
<td>SS</td>
<td>Reject</td>
<td>Nonnarrative</td>
<td>0.0</td>
</tr>
<tr>
<td>B-4</td>
<td>PS</td>
<td>Reject</td>
<td>Nonnarrative</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td></td>
<td>1.7</td>
</tr>
<tr>
<td>C-1</td>
<td>HU</td>
<td>Accept</td>
<td>Nonnarrative</td>
<td>3.7</td>
</tr>
<tr>
<td>C-2</td>
<td>SS</td>
<td>Accept</td>
<td>Nonnarrative</td>
<td>0.0</td>
</tr>
<tr>
<td>C-3</td>
<td>PS</td>
<td>Accept</td>
<td>Nonnarrative</td>
<td>1.6</td>
</tr>
<tr>
<td>C-4</td>
<td>BS</td>
<td>Accept</td>
<td>Nonnarrative</td>
<td>0.0</td>
</tr>
<tr>
<td>C-5</td>
<td>BS</td>
<td>Accept</td>
<td>Nonnarrative</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td></td>
<td>1.1</td>
</tr>
</tbody>
</table>

*HU = humanities, SS = social sciences, BS = biological sciences, PS = physical sciences.

Document Normalization as a General Filter Enhancement

The SourceFinder prototype evaluates feature values based on entire source documents. This includes front matter (e.g., abstracts and acknowledgements sections), as well as any other materials — such as footnotes, bibliographies, figures, and tables — that lie outside the body of a given source text. In contrast, when test developers examine potential sources, they presumably look primarily at the body of the text and not at front and back matter.

Failure to recognize the internal structure of a journal article in the electronic environment can obscure the role of individual text features, such as document length. For example, in computing length to meet copyright restrictions, presumably only the “body” of the document is relevant. Thus, a more precise implementation of the copyright length restriction
depends on document normalization. Currently, the SourceFinder feature *AlphaWordCount* (the length of the document in words) counts words everywhere, including bibliographies, which sometimes are rather lengthy. With a more abstract feature, such as *AvgWordDifficulty*, failure to ignore the text of footnotes and bibliographies would likely also skew these values.

To investigate potential filtering improvements associated with more accurate information about document length, a preliminary automated version of a document normalizer was developed. The normalizer was trained to automatically identify and exclude footnotes, tables, bibliography, and similar text. Table 16 shows contrasting data for five sample source documents that were analyzed before and after stripping off unwanted text. As the table shows, unwanted text was stripped two ways — by hand and using the normalizer described above. Two relevant features were evaluated for each document: the length of the document in words and the *UniqueWordperWord* feature. The *UniqueWordperWord* feature is used to detect documents containing a large number of unique words and is calculated by dividing the total number of unique words in a document by the total word count.

The data displayed in Table 16 indicate that SourceFinder examines an average of about 25% more text than needed, which necessarily skews all document features. For example, the average ratio of unique words to total words calculated from hand-stripped text is about .31, versus .22 for the entire document. These analyses indicate that a strategy of appropriately normalizing the text could lead to significant improvements in text categorization. However, additional analyses are needed to determine whether improved text classification can lead to needed reductions in the incidence of false-positive classifications.
Table 16

Feature Values of Full Text Versus Normalized Text

<table>
<thead>
<tr>
<th>Document</th>
<th>Content(^a) area</th>
<th>Normalization strategy</th>
<th>Length (words)</th>
<th>Unique words per total words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No stripping</td>
<td>5,267</td>
<td>.194</td>
</tr>
<tr>
<td>1</td>
<td>HU</td>
<td>Automated stripping</td>
<td>5,142</td>
<td>.241</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hand stripping</td>
<td>5,321</td>
<td>.300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No stripping</td>
<td>4,889</td>
<td>.276</td>
</tr>
<tr>
<td>2</td>
<td>HU</td>
<td>Automated stripping</td>
<td>3,601</td>
<td>.298</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hand stripping</td>
<td>3,660</td>
<td>.373</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No stripping</td>
<td>8,443</td>
<td>.173</td>
</tr>
<tr>
<td>3</td>
<td>BS</td>
<td>Automated stripping</td>
<td>5,374</td>
<td>.188</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hand stripping</td>
<td>5,729</td>
<td>.260</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No stripping</td>
<td>8,234</td>
<td>.194</td>
</tr>
<tr>
<td>4</td>
<td>BS</td>
<td>Automated stripping</td>
<td>6,051</td>
<td>.201</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hand stripping</td>
<td>6,177</td>
<td>.273</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No stripping</td>
<td>5,397</td>
<td>.212</td>
</tr>
<tr>
<td>5</td>
<td>PS</td>
<td>Automated stripping</td>
<td>3,626</td>
<td>.244</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hand stripping</td>
<td>3,448</td>
<td>.340</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>No stripping</td>
<td>6,446</td>
<td>.219</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Automated stripping</td>
<td>4,759</td>
<td>.309</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hand stripping</td>
<td>4,867</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) HU = humanities, SS = social sciences, BS = biological sciences, PS = physical sciences.

Analyses of Additional SourceFinder Features

An analysis was conducted to identify those SourceFinder features that are most highly correlated with human source acceptability judgments. A measure of human source acceptability was calculated for each document by summing the three numeric ratings obtained from the expert test developers. The resulting measure of human source acceptability ranged from a low of 3 (all three humans rated the document as definitely unacceptable) to a high of 15 (all three humans rated the document as definitely acceptable). The analysis indicated that two particular SourceFinder features have a relatively high correlation with human ratings of source acceptability: UniqueWordPerWord and KeyWordPerWord.

**UniqueWordPerWord**: As noted earlier, the SourceFinder feature labeled UniqWordPerWord is a measure of the size of the vocabulary used in a given document. This feature is also called type-token ratio because it is a ratio of the total number of word types (or unique words) in a document to the total number of word tokens in the document (Youmans,
The type-token ratio calculated for a particular document can range from 1, in the case of all unique words, to 1/N where N is the total number of words in the document. Certain word types, (e.g., words known to the information retrieval community as stop words, including “the” and “of”) occur relatively more frequently than so-called content words.

Table 17 summarizes the resulting correlations of the human source acceptability scores defined above with the *UniqueWordperWord* score calculated by SourceFinder. Coefficients are listed for the combined set of 79 documents (results for one document were unavailable at the time that these coefficients were being calculated) and for each of the four content areas. As can be seen, all of the estimated coefficients are negative. Thus, documents with a high density of unique words tended to be rated less favorably by test developers, especially in the humanities.

These results suggest that the *UniqueWordperWord* feature may help to detect documents that test developers rate as being inaccessible. Many of the documents that SourceFinder classified as acceptable were later deemed unacceptable by test developers because, as was noted in free-form comments, they tended to contain too much jargon or specialized vocabulary. This result was observed to occur most frequently for humanities documents. Thus, the current data support the hypothesis that SourceFinder’s *UniqueWordperWord* feature can help to detect documents with unacceptable levels of jargon or specialized vocabulary.

**Table 17**

**Correlation Results for SourceFinder’s *UniqueWordperWord* Feature**

<table>
<thead>
<tr>
<th>Content area</th>
<th>Number of documents</th>
<th>Correlation between the human acceptability score and SourceFinder’s <em>UniqueWordperWord</em> feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humanities</td>
<td>20</td>
<td>-0.59</td>
</tr>
<tr>
<td>Social sciences</td>
<td>19</td>
<td>-0.22</td>
</tr>
<tr>
<td>Biological sciences</td>
<td>20</td>
<td>-0.29</td>
</tr>
<tr>
<td>Physical sciences</td>
<td>20</td>
<td>-0.16</td>
</tr>
<tr>
<td>All documents</td>
<td>79</td>
<td>-0.29</td>
</tr>
</tbody>
</table>

*KeyWordperWord*. The SourceFinder feature labeled *KeyWordperWord* had the highest positive correlation with human judgments. This feature was developed as a measure of the degree of argumentation in a document. It is calculated by first counting certain key “argument”
words in a document and then dividing by the total number of words found in the document.

An analysis was conducted to determine whether the correlation of this feature with human source acceptability judgments was sensitive to the number of key terms included in the table of words classified as argument words. A total of 360 argument words were included in the original specification for this feature. As can be seen in Appendix B, which lists these 360 words, the original set of argument words included words that are only marginally related to level of argumentation (e.g., “cascading,” “event,” and “unfamiliar”). Based on previous results on the nature of “argument cue words” in discourse (cf. Schiffrin, 1980; Hirschberg & Litman, 1994; and Passonneau & Litman, 1997), this set was pruned to 66 words. A second, more drastic pruning resulted in a set containing only 14 argument words.

Table 18 displays the resulting correlations obtained with these three different key word lists. The results suggest that a strategy of using a smaller set of key words that is more narrowly tailored to the subjective notion of argumentation can result in enhanced filtering performance. In addition, our preliminary implementation of document normalization further boosted the correlations (Schapire, Singer, & Singhal, 1998).

Table 18
Correlation Results for SourceFinder’s KeyWordperWord Feature

<table>
<thead>
<tr>
<th>Size of key word table (in words)</th>
<th>Before document normalization</th>
<th>After document normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>360</td>
<td>0.29</td>
<td>0.32</td>
</tr>
<tr>
<td>66</td>
<td>0.35</td>
<td>0.42</td>
</tr>
<tr>
<td>14</td>
<td>0.34</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Analysis of Source-to-Passage Feature Variation

As noted earlier, the SourceFinder prototype’s feature acceptability thresholds were developed based on an analysis of selected, disclosed GRE reading comprehension passages. A comparison of two features — AvgSentenceLength and Type-Token Ratio, suggests that there are significant differences between the feature values of these GRE passages and the sources from which they were derived. Table 19 presents data on five such sources from which GRE passages were derived.

As the data indicate, there is a wide disparity between the feature values calculated in the
original sources and in the passages that were eventually developed from those sources. For example, the average sentence length of the five original sources is 23.2 words, compared with 35.0 words for the corresponding GRE passages. Similarly, the average type-token ratio for the five sources is .26, compared with .56 for the corresponding GRE passages. This analysis suggests that it may be possible to improve SourceFinder performance by using the newly developed corpus to derive new content-area threshold values for many of SourceFinder’s current features.

Table 19

*Feature Comparison of Source Documents and Corresponding GRE Passages*

<table>
<thead>
<tr>
<th>Document</th>
<th>Content&lt;sup&gt;a&lt;/sup&gt; area</th>
<th>Average sentence length (in words)</th>
<th>Type-token ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Original source</td>
<td>GRE passage</td>
</tr>
<tr>
<td>X-1</td>
<td>SS</td>
<td>20.1</td>
<td>32.3</td>
</tr>
<tr>
<td>X-2</td>
<td>SS</td>
<td>22.3</td>
<td>30.5</td>
</tr>
<tr>
<td>X-3</td>
<td>SS</td>
<td>24.0</td>
<td>29.0</td>
</tr>
<tr>
<td>X-4</td>
<td>BS</td>
<td>24.2</td>
<td>35.4</td>
</tr>
<tr>
<td>X-5</td>
<td>PS</td>
<td>25.4</td>
<td>47.5</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>23.2</td>
<td>35.0</td>
</tr>
</tbody>
</table>

<sup>a</sup> HU = humanities, SS = social sciences, BS = biological sciences, PS = physical sciences.

A second finding of this analysis is that for the 1,054 documents in the corpus, values for AvgSentenceLength range from 25 words to over 150 words per sentence. It is highly unusual for modern English sentences to be 100 or more words long, as an informal inspection of a typical document in the corpus quickly demonstrates. These preliminary findings illustrate further that the interpretation of an apparently straightforward notion such as average sentence length can be problematic.

*SourceFinder Interface Improvements*

It is important that the SourceFinder prototype system optimize interactions between the test developer and the Text Database. Figures presented earlier in this report provided samples of the original user interface. This study demonstrated that a modified interface must be able to
capture the following capabilities to make the SourceFinder system more useful and usable:

- **Content classification.** Test developers are assigned to create passages for a given content area. Screened documents in the SourceFinder database should be classified according to content categories to help limit the test developer’s search and save time.

- **Document length.** Test developers are assigned to create passages of a given length. Documents in the SourceFinder database should also be searchable by this feature.

- **Document browsing.** Test developers asked to be able to display the first few sentences of a candidate source document to enable them to decide if it warrants closer examination.

- **Document feature display.** Test developers asked for a prominent display of relevant source features with each candidate source.

- **Document review and tracking.** Test developers asked to be able to annotate the review status of candidate source documents online (e.g., reviewed-rejected, reviewed-accepted, not reviewed), to attach review notes to source documents, and to view the review status and other annotations made by others.

**Discussion**

**Corpus**

An important question in addressing the problem of locating usable source documents for GRE passages on the Web is determining the specific URLs to investigate. In this study, a corpus of 1,054 journal articles from the EBSCO Information Services database was collected, including documents SourceFinder would ordinarily reject. Each source was transferred to SourceFinder’s Text Database, where it is stored in a flat ASCII representation, along with values SourceFinder computed for 59 text features. This is a rich source of data from which to construct a profile of the URL that describes the usefulness of this site to GRE test developers. Test developers’ expert judgements of 80 documents accepted by SourceFinder and 20 rejected by the prototype tool are stored in this corpus. This information can be used to further train the SourceFinder search functionality. Researchers can use SourceFinder’s many features to analyze the document database, specify a more precise profile of a set of documents test developers might search through, and create more precise profiles of the partitions of this pool that are more or less acceptable to test developers.

The belief that acceptable sources for GRE reading comprehension items are rare has been reaffirmed by this study. Adding an enhancement to the SourceFinder interface to capture
test developers’ judgments will provide input to researchers for further improvements. Also, researchers will need to track SourceFinder’s database explorer search functionality when implemented to retrieve sources from Web sites. It is possible that these sites are not routinely updated and may have a fixed life cycle for use in GRE item sets.

**SourceFinder’s Filters**

SourceFinder’s filters are designed to work like a faucet. Implementing the filters with a loose set of thresholds is like turning the faucet on high. Because the loose constraints are easier to meet, a larger number of documents are allowed to pass through the filter and the “sink” fills up more quickly. Switching to a tight set of thresholds is like turning the faucet on low. Because the constraints are now more difficult to meet, fewer documents are allowed to pass through the filter and the sink fills up more slowly. If the filters were operating as intended, tight thresholds would result in more true positives and fewer false positives. However, the current study has demonstrated that a strategy of tightening SourceFinder’s feature thresholds does not result in a larger percentage of true positives. Thus, the study clearly demonstrates that improved filters are needed.

The greatest factor in improving SourceFinder’s filtering performance for locating GRE passages would be the ability to detect argument structure. This is an unsolved, nontrivial problem that requires extensive future research. Improvements to SourceFinder’s content and sensitivity filters are also needed. The current content filter is most effective for finding biological and physical science sources. The ability to detect sensitivity violations and to filter out unwanted aspects of documents (like bibliographies) would appear to significantly improve the SourceFinder prototype.

The findings of this study — particularly the data regarding the use of first person pronoun “I” to exclude first-person narratives and memoirs — suggest that more fine-grained features than those currently used are likely to enhance performance with greater reliability. SourceFinder currently computes the total proportion of pronouns in a document \( \text{PronounPerWord} \), using an unsorted set of English pronouns. By partitioning this rather heterogeneous set of pronouns into subsets that correspond to distinct narrative or stylistic functions, a feature that currently appears to be of limited value is likely to be enhanced.

Further, preliminary data suggest that the precise value of a useful threshold for such a first-person-pronoun filter differs depending on the content area of the document source. In
particular, the data suggest that humanities documents require a higher threshold than nonhumanities documents. Given that the style of scientific writing differs significantly in different scientific disciplines, it’s likely that several of SourceFinder’s stylistic features will require content-specific threshold values. Thus, additional subject-matter-specific analyses are warranted.

**System Performance**

This study evaluated SourceFinder’s system performance based on a strict test of its ability to perform a binary classification of documents (“accept” versus “reject”) as a stand-alone tool (without human intervention). Recall ranged from 38% to 87%, indicating that of the 16 documents deemed acceptable by human raters, SourceFinder retrieved a good proportion. Precision ranged from 13% to 17.5%, meaning that in the documents retrieved by SourceFinder, a large proportion were false positives. Thus, improved approaches for filtering out unacceptable documents are needed.

The performance metrics commonly applied to two-by-two agreement matrices (e.g., recall & precision) have pros and cons. One advantage is that a binary accept/reject classification is relatively noncommittal, and therefore, it is a good initial measure until it is possible to define more clearly how software such as SourceFinder should be integrated with the test development process. For example, should it operate as a stand-alone tool performing a simple binary classification? Will future analysis of human judgment data justify moving towards a more complex document classification? Should document location software allow for interactive searching?

**SourceFinder System Improvements**

A database explorer function needs to be added to the SourceFinder system. This will be accomplished through programming scripts that support customized Web crawling. The improvement will permit access to databases that contains abstracts and, in many cases, full-text articles, from a wide range of periodicals — including newspapers, scholarly journals, and specialized and general-interest magazines. The first SourceFinder Text Database was slow in response time due to the large amount of text retrieved. Future versions of SourceFinder will include a stateless database design that insures high performance and places less stress on the infrastructure.
Additionally, it is important that researchers have access to test developers’ judgements about source acceptability for future research. Specifications have been secured and programming has begun to enhance the interface. These new enhancements will allow users to (a) group source documents by relevant content categories, (b) display relevant information about each passage for easy reviewing, and (c) capture item writers’ reviews of sources. The latter enhancement will provide useful feedback to the research staff that can inform further SourceFinder improvements. The resulting data will also be useful for calculating interrater agreement. Figures 5 through 8 depict the newly designed SourceFinder user interface.

Note. Test developers can (1) recall previous search criteria and (2) set search criteria for content area, key words, and length.

Figure 5. New SourceFinder user interface: Search.
Note. Test developers can (1) view sources by content category and (2) read the first few lines of the source text.

**Figure 6.** New SourceFinder user interface: Browse sources.

**Figure 7.** New SourceFinder user interface: Review entire document.
Note. Test developers can (1) assign an acceptability rating, (2) select descriptive categories, and (3) provide comments about the source.

Figure 8. New SourceFinder user interface: Capture test developers’ judgments.

Directions for Future Research

This study yielded a promising methodology for evaluating source document selection that balances the quality of the results against the cost and effort of conducting the evaluation. For relatively little effort, data for a thorough evaluation were collected, including a database of 1,054 documents and human judgments for a subset of 100 of these documents. A great deal has been learned about the process test developers use to create GRE reading comprehension passages, and a foundation now exists on which to quantify improvements in successive generations of SourceFinder; this can be accomplished by re-evaluating the original set, as well as by evaluating entirely new document sets.
Corpus

Results presented above demonstrate that feature values can change significantly after document normalization. Preliminary findings conducted on six documents with and without materials that are not strictly part of the body of the text suggest that computation of values for such basic features as document length can be made more robust by focusing on methods to recognize the internal structure of a document in future work. Furthermore, for some features (e.g., any features dependent on word counts), it may be necessary to provide content-area specific options for various filtering processes. Scientific documents and the issues they raise in an electronic environment are quite distinct. For example, for scientific documents with high frequencies of numeric data, filters that can distinguish numeric “words” from other words, and intraword periods from sentence-boundary periods, presumably require distinct implementations from filters targeted towards humanities documents or a general robust implementation that would apply to both domains.

The low rate of acceptable articles found in the sample from the EBSCO Information Services database suggests numerous questions for future research to increase the efficiency of locating source documents. For example, further study of content-area differences is likely merited, in addition to further sampling to evaluate the validity of the rate of acceptance in our sample of 100 documents as a predictor for our corpus, or for other corpora constructed from this or other Web sites.

SourceFinder Features

Further analysis of the SourceFinder feature KeyWordperWord is currently being conducted. As described earlier, the KeyWordperWord feature computes a score based on a large set of key words. The hypothesis being examined is whether subsets of key words, developed through a more fine-grained analysis of the different types of arguments typically found in GRE passages, will result in improved performance. Preliminary results suggest that improved performance is indeed possible. For example, a strategy of using a shorter, more specific key word list increased the correlation with human acceptability ratings from 0.29 to 0.35. Our preliminary implementation of document normalization further boosted the correlation to 0.42. This suggests that further investigations designed to select optimal key word lists for different types of documents may help to enhance SourceFinder’s performance.

This study also demonstrated that documents in the four GRE content areas are not
homogeneous. Thus, a strategy of developing distinct document search heuristics for each subject area may also prove fruitful. Each of the analysis directions considered in this research (i.e., the development of targeted key word lists and the development of subject-specific search heuristics) should be addressed further in future research. Additional research is also needed to identify features for use in detecting acceptable sources in the women and minorities content area.

System Performance

Given this evaluation of the SourceFinder software, it is possible to attach baseline numeric values against which future iterations of the software, or the performance of other software on the set of documents with human judgments established here, can be compared. Success of future work could be defined as raising the value of precision without sacrificing recall.

The question of how to compare the behavior of an algorithm or software product to human performance on tasks that are inherently subjective is also an active area of investigation in the natural language processing community. Different methods of comparison, and different metrics to summarize the results of such comparisons, are being investigated (cf. the application of Cochran’s Q to human judgments of narrative event boundaries in Passonneau & Litman, 1997). Additional research is needed to determine the approaches that will be most useful for future SourceFinder evaluations. The projected model for integrating SourceFinder into the test development process also merits future reconsideration.
References


Notes

1 EBSCO Information Services provides several types of searchable databases. The Academic Search FullTEXT Elite host is a searchable database of abstracts, citations, summaries, and full text articles from over 4,000 periodicals, newspapers, magazines, and scholarly journals. A significant number of the periodicals that GRE and Graduate Management Admissions Test (GMAT®) item writers have used in print versions can be found here in electronic form.


3 “Current estimates are that it takes on average 7.5 hours for an item writer to identify usable source materials (experts are faster while it could take less experienced writers as much as three days)” (Bauer & Jha, 1999, p. 3).

4 Plans are now underway to enable SourceFinder to function automatically as a database explorer, so that its search component can be constrained to retrieve text from user-specified online databases, such as the repository of journal articles found in the EBSCO database.

5 Note that the process for calculating consensus judgments is similar to the adjudication procedure used in many constructed-response scoring applications. That is, final combined scores are calculated from only those observations remaining after the most discrepant scores have been removed.

6 Because the proportion of desirable documents in the sample is low, recall is very sensitive to different methods of defining the “human standard.” For example, if one averages the human judgments rather than using the consensus judgment, recall is .50. Note that precision remains the same, which is partly a reflection of the much greater proportion of documents that are relatively undesirable.

7 This simple method could potentially be refined to produce higher quality results with further corpus analysis (e.g., analysis of the contexts in which “I” occurs in documents from different subject matter areas).
## Appendix A: SourceFinder Feature List

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>AllAlphaWordCount</td>
<td>NonStopwordCount</td>
</tr>
<tr>
<td>AvgParaDifficulty</td>
<td>NonStopwordPerWord</td>
</tr>
<tr>
<td>AvgParaLength</td>
<td>NumwordCount</td>
</tr>
<tr>
<td>AvgSentenceDifficulty</td>
<td>NumwordPerWord</td>
</tr>
<tr>
<td>AvgSentenceLength</td>
<td>OneSyllWordCount</td>
</tr>
<tr>
<td>AvgWordDifficulty</td>
<td>OneSyllWordPerPara</td>
</tr>
<tr>
<td>AvgWordLength</td>
<td>OneSyllWordPerSentence</td>
</tr>
<tr>
<td>CharCount</td>
<td>OneSyllWordPerWord</td>
</tr>
<tr>
<td>DocDifficulty</td>
<td>ParaCount</td>
</tr>
<tr>
<td>DeterminerCount</td>
<td>PrepositionCount</td>
</tr>
<tr>
<td>DeterminerPerWord</td>
<td>PrepositionPerWord</td>
</tr>
<tr>
<td>HeaderCount</td>
<td>PronounCount</td>
</tr>
<tr>
<td>KeywordCount</td>
<td>PronounPerWord</td>
</tr>
<tr>
<td>HeaderPerPara</td>
<td>SentenceCount</td>
</tr>
<tr>
<td>KeywordPerPara</td>
<td>SentencePerPara</td>
</tr>
<tr>
<td>KeywordPerSentence</td>
<td>SixCharWordCount</td>
</tr>
<tr>
<td>KeywordPerWord</td>
<td>StopwordCount</td>
</tr>
<tr>
<td>MaxCharCount</td>
<td>StopwordPerWord</td>
</tr>
<tr>
<td>MaxKeywordCount</td>
<td>SyllableCount</td>
</tr>
<tr>
<td>MaxNonStopwordCount</td>
<td>ThreeSyllWordCount</td>
</tr>
<tr>
<td>MaxSentCharLength</td>
<td>ThreeSyllWordPerPara</td>
</tr>
<tr>
<td>MaxSentWordLength</td>
<td>ThreeSyllWordPerSentence</td>
</tr>
<tr>
<td>MaxParaCharLength</td>
<td>ThreeSyllWordPerWord</td>
</tr>
<tr>
<td>MaxParaWordLength</td>
<td>UniqWordCount</td>
</tr>
<tr>
<td>MaxParaSentLength</td>
<td>UniqWordPerWord</td>
</tr>
<tr>
<td>MaxStopwordCount</td>
<td>WordCount</td>
</tr>
<tr>
<td>MaxSyllableCount</td>
<td>WordDifficulty</td>
</tr>
<tr>
<td>MaxWordCount</td>
<td>WordPerPara</td>
</tr>
<tr>
<td>MaxWordDifficulty</td>
<td>WordPerSentence</td>
</tr>
</tbody>
</table>
Appendix B: SourceFinder Key Word List

“according”, // from: “according to”
“accurate”,
“accurately”,
“adapt”,
“adaptability”,
“adaptable”,
“adaptation”,
“adequate”,
“admittedly”,
“aggregate”,
“alternative”,
“alternatively”,
//”although”,
“analogous”,
“analogously”,
“analyses”,
“analysis”,
“analyze”,
“approach”,
“approximate”,
“approximately”,
“arbitrarily”,
“arbitrary”,
“arguably”,
“argue”,
“argument”
“assume”,
“assuming”,
“assumption”,
“balanced”,
“basis”,
//”because of”,
“because”,
“behavior”,
“belief”,
“bigger”,
“but”,
“cascading”,
“categorize”,
“cause”,
“characteristic”,
“commensurately”,
“compare”,
“comparison”,
“compatible”,
“competing”,
“complex”,
“complexity”,
“concede”,
“conceive”,
“conceivably”,
“conclude”,
“conclusion”,
“condition”,
“conditional”,
“conditionally”,
“consequence”,
“consequently”,
“consider”,
“consideration”,
“consistent”,
“consistently”,
“contrary”, // from: “contrary to”
“contrast”, // from: “in contrast”,
“controversy”,
“controversial”,
“controversially”,
“converse”,
“conversely”,
“convincing”,
“convincingly”,
“corollary”,
“corrective”,
“correspondingly”,
“counteract”,
“counterproductive”,
“critical”,
“criticism”,
“criticize”,
“cumulative”,
“decidedly”,
“deduce”,
“deduction”,
“define”,
“definition”,
“demonstrably”,
“demonstrate”,
“demonstration”,
“design”,
“despite”, “determination”,
“determine”,
“detrimental”,
“discernible”,
“discover”,
“discuss”,
“discussion”,
“disputable”,
“dispute”,
“distinct”,
“distinguish”,
“domain”,
“due”, // from: “due to”
“eliminate”,
“entail”,
“entity”,
“environment”,
“equally”,
“equivalence”,
“equivalent”,
“establish”,
“estimate”,
“evaluate”,
“evaluation”,
“event”,
“evidence”,
“examination”,
“examine”,
“except”,
“exception”,
“exhibit”,
“exhibition”,
“exist”,
“existential”,
“experience”,
“experiment”,
“explain”,
“explanation”,
“explicit”,
“exponential”,
“exponentially”,
“failure”,
“familiar”,

“feasible”,
//”for example”,
“footnote”,
“granted”,
“greater”,
“guarantee”,
“hence”,
“hypotheses”,
“hypothesis”,
“hypothesize”,
“hypothetical”,
“hypothetically”,
“idea”,
“identical”,
“illogical”,
“impact”,
“implication”,
“implicit”,
“implies”,
“imply”,
“incompatible”,
“inconsistent”,
“inconsistently”,
“increase”,
//”in addition to”,
//”in fact”,
//”in order to”,
//”in particular”,
//”in response to”,
//”in the guise of”,
//”indeed”,
“indicates”,
“indisputable”,
“indisputably”,
“induce”,
“induction”,
“inextricably”,
“infeasible”,
“infer”,
“inform”,
“insight”,
“insightful”,
“instability”,
//”instead”,

46
“relation”,
“relative”,
“relatively”,
“reliability”,
“require”,
“requirement”,
“requires”,
“research”,
“respective”,
“respectively”,
“result”,
“reveal”,
“rigorous”,
“rigorously”,
“riskier”,
“salient”,
“science”,
“severe”,
“severity”,
//”shows”,
“significantly”,
“similarly”,
“simultaneous”,
“simultaneously”,
//”since”,
“smaller”,
//”so”,
“specific”,
“specifically”,
“specified”,
“specify”,
“speculate”,
“speculative”,
“stable”,
“stability”,
//”still”,
“stipulate”,
“strengthen”,
“strictly”,
“structure”,
“structured”,
“study”,
“subsequent”,
“subsequently”,
“substantially”,
“success”,
“successive”,
“sufficient”,
“sufficiently”,
“suggest”,
“summarize”,
“suppose”,
“supposition”,
“susceptibility”,
“susceptible”,
“systemic”,
“tangible”,
“technique”,
“temporarily”,
“theoretical”,
“theories”,
“theory”,
“therefore”,
“thus”,
“tractable”,
“tradeoff”,
“traditional”,
“truth”,
“ultimately”,
“unconditionally”,
“unconditionally”,
“undeniably”,
“underestimate”,
“understand”,
“understanding”,
“undoubtedly”,
“unfamiliar”,
“uniformly”,
“unique”,
//”unless”,
“unmatched”,
“unquestionably”,
“unstable”,
//”unusual”,
“valid”,
“validate”,
“validity”,
“variance”,
“variation”,
“weaken”,
“weigh”,
“yield”
Appendix C: List of Journals Used for Study Source Documents

*African American Review* — social sciences (some humanities)
*African Journal of Economics and Sociology* — social sciences
*American Scholar* — humanities
*Art Journal* — humanities
*Astronomy* — physical sciences
*BioScience* — biological sciences
*Black American Literature Forum* — humanities
*business Horizons* — business
*Commentary* — social sciences
*Discover* — biological sciences and physical sciences
*Feminist Studies* — social sciences (some humanities)
*Journal of American Ethnic History* — social sciences
*Journal of Black Studies* — social sciences (some humanities)
*Journal of Social History* — social sciences
*The Nation* — social sciences
*Science News* — biological sciences and physical sciences
*Scientific American* — biological sciences and physical sciences
*Sloan Management Review* — business
*Smithsonian* — social sciences, biological sciences, and physical sciences
*Technology Review* — biological sciences and physical sciences
*Wilson Quarterly* — social sciences
*Women’s Review of Books* — humanities
Appendix D: Instructions to Raters

SourceFinder — GRE study
Instructions for Article Readers

Dear Readers:

The review of the articles will occur in a timed 2-stage format. Please also keep records of your time.

Stage 1 — Allow yourself a maximum of 1.5 hours per each content area for this stage.
• Start the clock.
• Look at each document and first decide if it is classified in the appropriate content area. If yes, write “Appropriate Content Area” on the page. If not, please write “Wrong Content Area” and also indicate which content area would be correct.
• Look at each document and decide “Is it a ‘definite accept,’ a ‘definite reject,’ or a ‘don’t know yet?’”
• Write “definite accept or reject” on the appropriate documents along with a quick, impressionistic note saying why, such as “clear arguments,” or “easy to understand but uninteresting,” or other thoughts.
• If the article is not a definite accept or reject, please write “don’t know yet” on the document and set in separate pile.
• At end of all 15 documents or 1.5 hours, whichever comes first, stop the clock.
• Count all the documents you’ve “processed.”
• Divide time elapsed by number of documents processed to get an estimate of “minutes per document.”

Stage 2 — Allow yourself maximum of 3 hours for this stage.
• Start the clock.
• Consider each document further in the “don’t know yet” pile and decide “Is it possible that this article could be a definite accept or reject?”
• If yes, write “probably accept or reject” on the document along with a quick, impressionistic note saying why, such as “very interesting, possibly a good argument” [accept] or “too technical” [reject].
• If no, write “still undecided” on the document along with a quick note saying why, such as “has potential, but too many threads in the argument to decide quickly.”
• At end of this pile of documents or 3 hours, whichever comes first, stop the clock.
• Divide time elapsed by number of documents processed to get an estimate of “minutes per document.”

Then please send all 20 documents (with decisions and notes written on them) back to Lisa Hemat at Mail Stop 11-R by (Jan 19?). Also, please let us know the time estimates and numbers of documents processed for each stage.
Appendix E: Sample Documents That Illustrate Eight Categories

1. A sample document that was classified as having the desired level of argumentation. Words and phrases that appear to be markers of the document’s argument structure are shown in boldface.

**Document ID = Social Science 5A**
LAST YEAR, Americans donated over $190 billion to charity, gave roughly 20 billion hours of their time as volunteers, and participated in nearly 2 million tax-exempt organizations, not to mention the still greater number of less formal groups through which they performed an astonishing range of public-spirited works. Indeed, except for a minor dip in the early 1990’s, Americans have been giving and volunteering at an ever-increasing rate for over twenty years.

Yet despite this apparent vitality, all is far from well with civil society in America—or at least so argues the Harvard political scientist Robert D. Putnam in his much-talked-about new book, Bowling Alone: The Collapse and Revival of American Community.(A) By Putnam’s lights, the United States is in fact suffering from a severe breakdown of civic and social ties, a development that is partly responsible not only for the public’s heightened mistrust of government but even for the growing incidence of depression and suicide. Nor is this withdrawal from civic engagement something new: with the exception of older Americans, Putnam claims, it has been going on for more than 30 years.

How can two such different pictures be reconciled?
IT WAS the 19th-century French observer Alexis de Tocqueville who first suggested that voluntary associations in America had a value that went well beyond their own immediate goals. Such groups, he wrote, by taking on tasks that would otherwise have to be performed by government, kept the state from growing too powerful. More importantly, they taught those who participated in them both the norms of community life and the skills of cooperation necessary for sustaining a democratic society. In a phrase coined by the distinguished sociologist James S. Coleman, who devoted much of his career to developing Tocqueville’s insight, associations helped people to acquire “social capital,” a commodity every bit as valuable for getting ahead, Coleman maintained, as its more familiar financial counterpart.

Himself a strong proponent of this view, Putnam first sounded an alarm about the erosion of social capital in the United States in a 1995 article in the Journal of Democracy. Since the mid-1960’s, he reported with distress, the membership rolls of a great variety of long-established groups--religious congregations, civic and fraternal orders, labor unions, the Boy Scouts, the Red Cross, parent-teacher associations (PTA’s)--had steadily fallen. And as for the organizations whose ranks had actually grown during this time, many of them, like the American Association of Retired Persons, demanded little more involvement than the writing of a check. Even the country’s less formal venues for socializing seemed threatened, Putnam suggested. The whimsical title of his article, “Bowling Alone,” came from his discovery that fewer people were joining bowling leagues, a matter of no small concern to bowling-alley proprietors whose revenues depend heavily on the convivial sharing of beer and pretzels.

The article was an intellectual sensation, propelling its author to meetings with President Clinton and appearances on television and in the popular press. But it also attracted a stream of critics, most of whom took issue with Putnam’s findings.
Untangling the biology of the hair follicle
From the biblical tale of Samson and Delilah and the fairy tale of Rapunzel to the eponymous 1960s rock musical and the latest coiffures, hair has long been a source of fascination and obsession. Both too little hair and too much hair can cause psychological and social anguish. And the multibillion dollar industry devoted to products designed to alter hair growth and appearance attests to the importance that humans place on healthy and attractive hair.

Aside from its importance in social interactions, hair serves a number of purposes for humans and other mammals. It provides insulation, acts as a sensory organ, and helps disperse perspiration and pheromones produced by glands in the skin. Specialized immune cells at the opening of the human hair follicle can detect pathogens at the skin’s surface and activate an immune response. The follicle also contains epithelial stem cells that are essential for regenerating the follicle and are also involved in regenerating the epidermis, or upper layer of skin, when it is injured or abraded.

For over a century, biologists have been working to understand the structure and function of the hair follicle and the factors that control this tiny but complex organ’s growth and development. Recent progress in hair research, including advances made through molecular and genetic approaches, is revealing some of the intricacies of hair follicle biology and may eventually lead to the development of more effective treatments for hair growth disorders. The hair follicle is also attracting interest as a useful model system for studying a range of biological processes, including tissue development, epithelial cell differentiation, apoptosis (programmed cell death), and tumor formation.

An introduction to hair
Like fingernails, claws, feathers, and scales, hair is a specialized skin appendage. Each hair follicle is associated with a sebaceous (oil) gland that opens to the skin through the follicle. The mature hair follicle (see diagram page 305) reaches down from the epidermis into the lower layer of skin, or dermis, a dense connective tissue that is richly supplied with blood vessels and nerves. The hair shaft, or fiber, that is produced by the hair follicle consists of two or three concentric layers and is composed largely of proteins called keratins. These tough, fibrous proteins are produced by rapidly proliferating matrix cells in the hair bulb, the lowermost portion of the follicle. Melanocytes in the hair bulb produce the pigment, or melanin, that colors the hair shaft.

In the mature follicle, three concentric layers of epithelial cells surround the hair shaft, forming the inner root sheath. The outer root sheath, which is continuous with the epidermis and the sebaceous gland, forms the outermost epithelial layer of the follicle. A specialized region of the outer root sheath, called the “bulge,” which is located just below the sebaceous gland, contains epithelial stem cells. Attached near the bulge region is the arrector pili muscle, which makes the hair stand erect in response to cold and other stimuli. A thin dermal sheath encloses the epidermal components of the follicle and is connected to the dermal papilla, a small, pear-shaped cluster of cells that is encircled by the hair bulb and plays a central role in hair follicle growth and development.
3. A sample document that was classified by SourceFinder and by the human raters as being the appropriate genre.

**Document ID = Biological Science 5A**

**NATURAL HISTORY OF REPTILIAN DEVELOPMENT: CONSTRAINTS ON THE EVOLUTION OF VIVIPARITY**

A powerful approach is clarifying the evolution of viviparity: This approach involves the identification and critical evaluation of closely related taxa that vary in reproductive mode.

Oviparity, or egg-laying, is the dominant mode of reproduction among vertebrates. Nevertheless, viviparity, the retention of the egg within the reproductive tract until embryonic development is complete, characterizes almost all mammals; it has also had at least 150 independent origins within the fishes, amphibians, and reptiles (Shine 1985, Blackburn 1992, Wourms and Lombardi 1992). These multiple origins suggest pervasive benefits to viviparity across a wide range of taxa, life histories, and habitats. In the squamate reptiles (lizards and snakes), for example, viviparity is the most common reproductive mode in cold climates, and recent origins of viviparity in this group are also associated with cold climates (Shine 1985). Gravid females in cold climates can thermoregulate to keep embryos warmer than they would be in a nest, thus enhancing development. Thermo regulation by the female may thus ensure that birth occurs at the appropriate season or even that reproduction is successful at all. Viviparity is also advantageous in very wet or dry habitats, for example, because it obviates the need for females to find suitable sites in which to lay their eggs.

In reptiles, viviparity is associated with a plethora of integrated morphological and physiological features that are not present in oviparous reptiles; these features are presumed necessary for successful embryonic development in the oviduct (Packard et al. 1977, Guillette 1993). Early insights into the evolution of these reproductive features were based on comparisons between typical oviparous and viviparous species. Some of the distinguishing features of the viviparous species examined were the major reduction or absence of an eggshell and the presence of some form of placentation (Weekes 1935 and included references). However, because the species used in these comparisons represented the extremes of a putative evolutionary sequence, their use as a model for elucidating the actual sequence or timing of the morphological and physiological changes attending the evolution of viviparity is limited. In fact, these observations are consistent with both a saltation model that posits that the characteristic features of viviparity arise suddenly and simultaneously, and a gradualist model that posits incremental evolution from one reproductive mode to the other (Blackburn 1992, 1995).

A relatively new and more powerful approach is clarifying the evolution of viviparity; this approach involves the identification and critical evaluation of closely related taxa that vary in reproductive mode (Guillette 1982, Heulin 1990, Mathies and Andrews 1995, Qualls 1996, Smith and Shine 1997, Mendez-de la Cruz et al. 1998). Of these studies, Qualls (1996) provides the best evidence that a gradual process describes the evolution of viviparity. His study on Lerista bougainvillii, an Australian skink, involved three conspecific populations: one oviparous, another viviparous, and a third that is morphologically and physiologically intermediate. An increase in the length of egg retention associated with reduction in eggshell thickness among these populations supports the hypothesis that viviparity evolves gradually from oviparity. The other studies presented data that are consistent with this interpretation—that is, conspecific populations or closely related species exhibited the expected grade of features intermediate between oviparity and viviparity.
4. A sample document that was classified by human raters as being inappropriate genre. This humanities article was classified by SourceFinder as physical sciences (probably due to the presence of the bolded earth-science terms.)

Document ID = Physical Science 12C
THE POETIC VISION OF SPANISH SCULPTOR EDUARDO CHILLIDA
EDUARDO CHILLIDA, THE RENOWNED 76-YEAR-OLD SPANISH sculptor, wants to climax a long and distinguished career by carving out a massive space stories nine high and just as wide inside a mountain on one of Spain’s Canary Islands in the Atlantic Ocean. The tall and soft-spoken Chillida, who often sounds more like a poet than a sculptor, is awed by the idea of standing within the enormous emptiness of a mountain and looking upward at shafts of light from the sun and the moon.

Chillida (pronounced Chee-YEE-dah) may never realize the work. Although the provincial government of the Canary Islands has approved the project, and promoters are already urging tourists to visit the anointed mountain, a small group of environmentalists has denounced the venture, castigating Chillida for meddling with nature. On top of this, engineers have not yet finished a study to determine whether Chillida’s plan is structurally sound, and other problems have arisen. Whether successful or not, the grand ambition of the mountain project has not surprised anyone who knows the work of Chillida well. The artist, who has created both monumental and smaller pieces out of iron and steel and wood and alabaster and cement and day and paper and stone and plaster, has come to look on space itself as material to mold. To sculpt an immense, mind-boggling space inside a mountain makes artistic sense to him.

Tindaya seemed ideal. A mining company was already extracting trachyte, a decorative rock, from one side of the mountain, and Chillida could position the entrance to the interior on the site of the quarry. As Chillida envisioned his mountain sculpture, a visitor would stand at the entrance and look back at more than a mile of lava-laden earth extending to the sea. The visitor would then walk down a corridor almost the length of a football field before reaching the massive cube of interior space carved within the mountain. Light would come from two shafts that would penetrate the summit. The visitor, Chillida hopes, would feel the smallness of mankind and the immensity and force of space.

But Chillida did not anticipate the problems that would be spawned by his project. Delays have been caused by environmentalists afraid of damage to the mountain, anthropologists worried about the loss of what they believe may be the footprints of prehistoric man, engineers who have still not completed a study on whether the digging can be done safely, and investigators looking into accusations of corruption between politicians and the mining company that has been quarrying the trachyte. Although the Canary Island government gave the project final approval in late 1998 Chillida is concerned that politicians might back down if the problems intensify. The frustration has made him testy and defiant. He told reporters recently, “The Tindaya project will be finished someday even if I am not here to see it.”

The troubles over the mountain coincided with a bout of near depression. But Chillida, in an unusually frank interview not long ago, did not blame the controversy for the attacks that made him feel like “everything was falling apart.” The sculptor conquered this funk by listening to Bach and other music, reading philosophy and poetry, and forcing himself by sheer will to act as if he did not feel the near depression. Still a productive artist these days, Chillida is powered by the sentiment expressed in the line of a poem that he has cited often over the years. Quoting the French poet Rene Char, Chillida has told many inter viewers trying to understand his work, “One must walk forward into the night.”
After a decade of calculations, the first wave of materials designed from scratch on the computer are ready to be made and tested. On the horizon: new substrates for optics and Electronics. The first thing you notice about Gerbrand Ceder’s materials science lab at MIT is that there are no crucibles, no furnaces, no crystal-growing instruments. Instead, you find a row of high-resolution computer displays with grad students and postdocs tweaking code and constructing colorful 3-D images. It’s in this room, quiet except for the hum of fans cooling the computer power, where new high-tech ceramics and electronic materials that have never been seen or made before are being forged. They are taking form “in virtuo”--designed from scratch on the computer, distilled out of the basic laws of physics.

The next thing you’re likely to notice is how young Ceder is. Quick to laugh but intensely passionate in explaining his work, the 33-year-old associate professor is one of a new breed of materials researchers, trained in traditional processing techniques, who have turned to discovering materials using computers. The dream is simple: Replace the age-old practice of finding new substances by trial and error, with calculations based on the laws of quantum mechanics that predict the properties of materials before you make them.

You can, in theory at least, design metals, semiconductors and ceramics atom by atom, adjusting the structure as you go to achieve desired effects. That should make it possible to come up with, say, a new composition for an electronic material much faster. Even more important, tinkering with atomic structure on a computer makes it possible to invent classes of materials that defy the instincts of the trial-and-error traditionalists.

It’s an idea that has been kicking around for at least a decade. But with the explosion in accessible computer power, as well as the development of better software and theories, it’s becoming a reality. Last year, Ceder and his collaborators at MIT synthesized one of the first materials that had actually been predicted on a computer before it existed. This new aluminum oxide is a cheap and efficient electrode for batteries. And while it may or may not lead to a better, lighter rechargeable battery, the success of Ceder’s group--and related work at a handful of other labs--is proving that useful materials can be designed from the basic laws of physics.

Designing from first principles represents a whole new way of doing materials science, a discipline that Ceder describes as “a collection of facts with some brilliant insights thrown in.” It’s a transformation he’s been aiming at since his undergraduate days in the late 1980s at Universite Catholique de Louvain in Belgium. “My background is heat and beat metallurgy,” he explains. “But I always thought there should be more to it, some way to calculate things using all the great physics of quantum mechanics.”

Getting there, however, won’t be easy. Scientists have known for decades that, according to the rules of quantum mechanics, if you could detail the position of the electrons swarming around atoms, you could then calculate physical properties of the material. Yet the sheer difficulty of carrying out these calculations has made the task seem hopeless. The computations are hard for even one molecule, but for the huge numbers of atoms that make up even the smallest chunk of a solid material, the chore is truly intimidating.
SALLY’S RAPE: ROBBIE MCCAULEY’S SURVIVAL ART

In the climactic scene of Sally’s Rape, African-American performance artist Robbie McCauley stands naked on an auction block, encouraging spectators to bid on her body, while she describes the sale and repeated sexual abuse of her great-great-grandmother, a slave.[sup1] As several feminist performance theorists have noted, this particularly vivid image of McCauley crystallizes key issues in our discourse, such as the display of the black female body, narratives of historical revision, and the centrality of identity, despite its various contingencies.[sup2] In this scene of bodily spectacle, as in her more subtly crafted dialogue, how does McCauley manage to reclaim her body from the inscriptions which have persistently haunted representations of women of color: the exotic other, white-man’s pawn, tragic victim? Using black cultural studies and feminist performance theory, I will discuss how McCauley creates a space for self-representation, for emotional and intellectual reflection on a painful past, for talking back to the history of victimization, and dismantling the structures of stereotype.[sup3]

Sally’s Rape is a social experiment in which Robbie McCauley, an African-American female performance artist, performs the black female subject out of victimization. Like any social or theatrical experiment, the results are rather inconceivable to gauge. However, according to my own reception, and that of other spectators, my evaluation is optimistic. McCauley’s contribution to the emerging black female theatrical subject is her development of an anti-racist, heuristic performance mode.[sup1] She inherits a tradition of black performance which is both politically and mimetically sophisticated, expanding it to express the often obscured experience of gender. McCauley’s performance experiments demonstrate a black female subject bearing witness to the confluent demons of racism and sexism in representation as well as in everyday life. In this essay, I will explicate McCauley’s key heuristic tools--revision, embodiment, and dialogue--in the performance text of Sally’s Rape.

Sally’s Rape shares the theme of survival with two other performance pieces, usually grouped under the series title “Confessions of a Working Class Black Woman.” Since the mid-1980s, McCauley has performed this series as works-in-progress, all of which center on stories from her family history. The first, My Father and the Wars, concerns McCauley’s relationship with her father, and his life in military service. Indian Blood, part two, focuses on her Native-American grandfather’s participation in the genocide of his own people. In the third piece, Sally’s Rape, McCauley shifts her focus to the experiences of women in her family. Each performance is about an ancestor’s survival, but also about how McCauley tells their stories in painfully acute enactments which demonstrate the surviving impact of past events on present racial conflicts.[sup4]

Sally’s Rape: Stories, Enactments, Conversation

Describing Sally’s Rape is difficult, not only because of the intensity of the material but also because the performance text has varied greatly over the course of several years. It is now available in an anthology of plays by African-American women, but this published version was transcribed from a single event and cannot represent the many variations of this work-in-progress. Its inclusion in an anthology is important, however, because it will allow the play to reach a much wider audience, offering a powerful representation of the black female subject in an interrogation of American culture.
ETHNIC SOLIDARITY AND BLACK BUSINESS: THE CASE OF ETHNIC BEAUTY AIDS DISTRIBUTORS IN CHICAGO

ABSTRACT. This paper examines the development of a separate economic niche for black entrepreneurs in Chicago’s ethnic beauty aids industry. It argues that this economic niche developed in response to advantages black entrepreneurs had in mobilizing ethnic resources in the black community. The paper’s findings lend support to general theoretical arguments stating that ethnicity, race, and other symbols for identity function as low-cost screening devices for evaluating the likelihood that trading partners will honor economic contracts, particularly in a market setting where formal market mechanisms are not fully developed. The findings are based on a series of in-person interviews with Korean, Jewish and black distributors of ethnic beauty aids in Chicago. Although this paper focuses on a single market niche, its conclusions indicate that greater attention needs to be paid to the effects of social, political, economic and structural factors on minority business development.

Introduction

Currently, the relative importance of class and ethnic resources in minority business development has become the focus of debate among scholars interested in the study of entrepreneurship. For instance, Bates (1998) points out that, in general, minority business development is shaped by relative advantages that minority entrepreneurs have in mobilizing financial and human capital. Consequently, he argues that differences in the levels of entrepreneurial activity between black Americans and Asian immigrants primarily result from the varying ability of entrepreneurs in each group to mobilize class resources, and not from factors linked to either group’s access to respective ethnic networks. However, other scholars have argued that ethnic resources do play an important role in the entrepreneurial process, particularly in settings where capital constraints exist and where formal market mechanisms are underdeveloped. For instance, Landa (1994, 28) argues that, “under conditions of contract uncertainty, where the legal framework for the enforcement of contracts is not well developed, the identity of a potential trading partner matters.” In such settings, “kinship, ethnic, and religious stature as well as other symbols of identity...serve as low-cost signaling and screening devices that allow a trader to choose to trade with only those traders who one perceives to be trustworthy or reliable in honoring contracts” (Landa 1994, 29).

In general, Bates (1998) makes an important point concerning the link between financial resources, human capital, and successful entrepreneurship. However, these factors seem to be more predictive of successful entrepreneurship when considered within the context of well developed markets. In less formalized market settings, it is hypothesized that ethnic or group resources have a stronger influence on entrepreneurial activity. In this paper, this hypothesis concerning the relationship between market structure and ethnic resource mobilization will be examined within the context of black business development in the ethnic beauty aids industry.

The ethnic beauty aids industry specializes in the manufacturing, distribution and sale of hair care products and cosmetics for black consumers. As a result, this industry focuses on
developing consumer markets almost exclusively in the black community. The ethnic beauty aids industry is divided into two distinct economic niches. One niche focuses on retail sales in the black community. This niche is embedded in a highly formalized market setting, and entrepreneurial success is strongly influenced by relative advantages that entrepreneurs have in mobilizing financial resources and human capital. In fact, despite their placement in a market context that provides goods and services to a predominantly black clientele, Korean and Jewish entrepreneurs maintain a high profile in the retail niche of the ethnic beauty aids industry, while black entrepreneurs are rarely found in this market niche. However, black entrepreneurs are highly visible in the other niche of the ethnic beauty aids industry. This niche focuses on the sale of professional products to black beauticians and barbers. Unlike the retail niche, the professional niche of the ethnic beauty aids industry is smaller, less formalized, and driven by more intense social interactions between economic actors.

8. A sample document that was classified as containing words and content that would require background knowledge in order to complete successfully.

Document ID = Social Science 15C
BEYOND TWIN DEFICITS: EMOTIONS OF THE FUTURE IN THE ORGANIZATIONS OF MONEY
ABSTRACT. This paper outlines new developments in the sociology of money. It highlights certain aspects of Post Keynesian monetarism and explores Keynesian concepts of emotions relative to economics and economic sociology. Gunnar Myrdal’s work on time and money contributes to the discussion. Underdeveloped areas of discourse in both sociology and economics are identified and the resulting superficiality of references to money are examined. Sociology, for example, has historically neglected concepts of future time and money, while economics has paid little attention to emotions and organizations. Removing these orthodox barriers allows economics to be informed by concepts previously relegated to sociology, such as emotions of trust and confidence. This process may induce the disaffected from both disciplines to draw from each other, creating an alternative, and ultimately more satisfactory understanding of money.

Introduction
MONEY AND HIGH FINANCE ARE MAJOR CONTEMPORARY CONCERNS. Advocates of market orthodoxy are having difficulty advancing the merits of further financial liberalization to an ‘informed public’, and even former proponents are expressing disenchantment.(n1) One need only consider the failure of ‘shock therapy’ for Russia, where a formal finance structure was lacking; or the World Bank’s criticisms of the IMF’s policy toward Indonesia in 1998, where unemployment, distress and military violence mounting daily.(n2) The heartland of world finance, Wall Street, attempting to suggest that the 1997-1998 Asian financial crises were only due to ‘cronyism’, was startled in late 1998 by the sudden collapse of a US hedge fund ironically called Long Term Capital Management (LCTM). It was alleged to have borrowed US$1.25 trillion, investing twenty times its own capital on some holdings. The large investment banks (e.g. Merrill Lynch and Goldman Sachs), required by the Federal Reserve to rescue it, had been kept in the dark about LCTM’s long-term exposure; apparently influenced by the financial sector’s blind faith in LCTM’s so-called ‘gurus’ John Meriwether and two Nobel-winning economists.(n3) In other words, 1998 was not a good year for high finance.
Amidst these crises, critiques of market liberalism are a ‘growth industry’. New arguments for international and national regulation are important but few offer assessments of the limits of proposed remedies. More necessary than criticism, however well-justified, are innovative theoretical challenges, which will engender alternative research programs and alternative policy debates. Contemporary problems are too pressing for analysis to remain constrained by disciplinary boundaries. Sociology has been reluctant to assess the financial side of economic life with its intensely future orientation. In the same way, economics has been relatively uninterested in the hierarchical, unequal distribution of power and the conflicts (not mere competition) for control over opportunities that largely take organizational forms.

Where might a fruitful meeting place on money exist? Economists and sociologists agree that a money economy cannot exist without a normative consensus and trust, backed by state authorities (Mizruchi & Stearns 1994 p. 317). Trust is central to the existence of money, however most monetary theories treat it as an ad hocery (Ingham, 1996a p. 250). In contrast, sociological explorations of trust are well-developed and linked to the concept of risk (e.g. Niklas Luhmann, Ulrich Beck). In the case of finance organizations, trust and risk are pre-eminent but rarely defined in these sociological terms (despite their reassuring names, such as Bankers Trust). Thus, where economics and sociology intersect, sociology is capable of analyzing trust but needs active, cross-disciplinary discussions with economists about money. Economists, in contrast, may be interested in the converse.