Listening. Learning. Leading. ™

Supporting Efficient, Evidence-Centered Item Development for the GRE Verbal Measure

Kathleen M. Sheehan

Irene Kostin

Yoko Futagi

July 2007

ETS GRE Board Research Report No. 03-14

ETS RR-07-29

Supporting Efficient, Evidence-Centered Item Development for the GRE Verbal Measure

Kathleen M. Sheehan, Irene Kostin, and Yoko Futagi

ETS, Princeton, NJ

GRE Board Research Report No. 03-14

ETS RR-07-29

July 2007

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

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 no necessarily represent official Graduate Record Examinations Board position or policy.

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

****

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

Educational Testing Service
Princeton, NJ 08541

Abstract

This paper explores alternative approaches for facilitating efficient, evidence-centered item development for a new type of verbal reasoning item developed for use on the GRE® General Test. Results obtained in two separate studies are reported. The first study documented the development and validation of a fully automated approach for locating the types of stimulus materials needed to support more efficient generation of the studied item type. The validation results confirmed that the proposed filtering technique can help item writers increase the percentage of acceptable stimulus paragraphs located per unit time interval from the current level of about 10% to nearly 30%. The second study documented the development and validation of a set of task models designed to help item writers generate new items that are optimally constructed to provide high-quality evidence about targeted skills. The proposed models were validated by considering the percentage of difficulty variance accounted for by the specified item classifications. That amount ranged from slightly more than 30% for items designed to test vocabulary skills to slightly more than 40% for items designed to test additional verbal reasoning skills, such as generating near and far inferences and understanding complex oppositional reasoning.

Key words: Evidence-centered design, ECD, item difficulty modeling, latent semantic analysis, LSA, targeted item writing, tree-based regression, the SourceFinder system, verbal reasoning skills, vocabulary knowledge

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

New test delivery technologies, such as Internet-based testing, have created a demand for higher capacity item generation techniques that are (a) grounded in a credible theory of domain proficiency, and (b) aligned with targeted difficulty specifications. This paper describes a set of automated text analysis tools designed to help test developers more efficiently achieve these goals. The tools are applied to the problem of generating a new type of verbal reasoning item called the paragraph reading (PR) item. This new item type was developed for use on the Graduate Record Examinations® (GRE®) General Test, an examination taken by students seeking admission to graduate school. It consists of a short passage, typically between 90 and 130 words, followed by two, three, or four items designed to elicit evidence about an examinee's ability to understand and critique complex verbal arguments such as those that are typically presented in scholarly articles targeted at professional researchers. This new item type was developed at ETS as part of an on-going effort to enhance the validity, security and efficiency of item development procedures for the GRE.

Two different approaches for enhancing the efficiency of the PR item development process are considered in this paper. The first approach focuses on the *passage development* side of the item writing task; the second approach focuses on the *item development* side of that task.

The approach for enhancing GRE passage development efficiency builds on previous research documented in Sheehan, Kostin, Futagi, Hemat, and Zuckerman (2006) and Passonneau, Hemat, Plante, and Sheehan (2002). This research was designed to capitalize on the fact that, unlike some testing programs that employ stimulus passages written from scratch, all of the passages appearing on the GRE verbal measure have been adapted from previously published source texts extracted from scholarly journals or magazines. Consequently, in both Sheehan, Kostin, Futagi, et al. (2006) and Passonneau et al. (2002), the problem of helping item writers develop new passages more efficiently is viewed as a problem in automated text categorization.

Automated text categorization techniques have reflected a renewed interest in recent years, due, in part, to the increasing number of documents now available in electronic form and the corresponding demand for more efficient organization and retrieval mechanisms. Recent areas of application have included automated genre-classification studies (e.g., Biber, 1988; Biber et al., 2004; Reppen, 2001), automated essay-scoring systems (e.g., Burstein, 2003; Larkey, 1998), and the automated source-analysis studies documented in Passonneau et al.

(2002) and Sheehan, Kostin, Futagi, et al. (2006). These latter two studies documented the development and validation of an automated text analysis system designed to help test developers find needed stimulus materials more quickly. The resulting system, called SourceFinder, includes three main components: (a) a database of candidate source documents downloaded from appropriately targeted online journals and magazines, (b) a source evaluation module that assigns a vector of acceptability probabilities to each document in the database, and (c) a capability for efficiently searching the database so that users (i.e., item writers) can restrict their attention to only those documents that have been rated as having a relatively high probability of being acceptable for use in the particular source-finding assignment at hand.

A detailed evaluation of the existing SourceFinder module is presented in Sheehan, Kostin, Futagi, Hemat, et al. (2006). That research confirmed that the existing module has contributed to a significant decrease in the time needed to find acceptable source material for the types of stimulus passages included on the current GRE verbal section. Since the new PR passages differ from the existing passages in a number of important ways, however, this study considered the problem of developing an additional SourceFinder module designed to help GRE test developers search for stimulus material appropriate for use in developing new PR passages and items.

As was noted above, this study also considers approaches for enhancing the efficiency of the PR item development process. Techniques for facilitating item development efficiency have been discussed by Sheehan and colleagues (Sheehan, 1997, 2003; Sheehan, Kostin, Futagi, et al, 2006; Sheehan & Mislevy, 1990). This research demonstrated that item writers can work more efficiently by generating new items that conform to prespecified task models designed to provide unambiguous evidence about examinees' mastery status on targeted proficiencies.

The remainder of this paper is organized into three sections. The first section summarizes the analyses implemented to develop and validate a new SourceFinder module designed to help test developers find acceptable PR source documents more quickly. The second section documents the analyses implemented to clarify the constructs assessed by the new PR item type and to develop task models that provide high-quality evidence about targeted skills. Finally, the third section presents conclusions, recommendations, and directions for future research.

Study 1: Enhancing Passage Development Efficiency via Automated Evaluation of Candidate Source Paragraphs

The SourceFinder database currently includes over 90,000 documents downloaded from over 60 different journals and magazines designated as potentially appropriate for use in developing new GRE passages and items. Estimates of the acceptability status of each document, relative to a specified number of potential passage development assignments, are stored along with each document. These estimates enable item writers to limit their search to only those documents that have been rated as having a relatively high probability of being acceptable for use in satisfying the particular passage development assignment at hand. Since PR passages are developed from paragraphs, as opposed to entire documents, however, the goal of this study was to assign a PR-specific acceptability rating to each paragraph in the database. Test developers can then use these new estimates, in combination with SourceFinder's existing search capability, to restrict their attention to only those paragraphs that have been rated as having a relatively high probability of being acceptable for use in developing a new PR passage.

Method

The methodology developed to obtain a PR-specific acceptability rating for each paragraph in the SourceFinder database included four stages: (a) data collection, (b) feature development, (c) dimensionality analysis, and (d) model development and validation. The analyses implemented at each stage are described below.

Stage 1: Data Collection

Training and validation datasets were assembled as follows. First, an initial training sample was assembled by randomly selecting 114 paragraphs from the SourceFinder database. The selected paragraphs were then presented to two GRE test developers for evaluation. Raters were asked to provide two types of ratings: (a) a quantitative estimate of each paragraph's "acceptability" status expressed on a 1–5 scale, where 1 = *definitely reject*, 2 = *probably reject*, 3 = *uncertain*, 4 = *probably accept,* and 5 = *definitely accept*, and (b) a brief, written description of the aspects of text variation considered during the evaluation process. Second, because the sample was not expected to yield a large number of acceptable paragraphs (i.e., paragraphs rated as probably or definitely accept), a supplemental sample of 47 *historical* paragraphs was also collected. Historical paragraphs are paragraphs that had been used previously to create

operational PR passages and items. This strategy yielded 47 additional training paragraphs classified at the definitely accept level and increased the size of the training sample to a total of 161 paragraphs.

An independent cross-validation data set was assembled. This data set included 1,000 additional paragraphs that each had been evaluated by two GRE test developers as part of their operational passage development work.[1] As was the case for the training sample, paragraph acceptability scores were specified on the 5-point acceptability scale.

Stage 2: Feature Development

Feature development was implemented by first examining the ratings and comments collected at Stage 1. Then, hypotheses were generated about the aspects of text variation that might account for the observed similarities in the comments provided for similarly rated paragraphs. Finally, natural language processing tools were developed to automatically extract candidate explanatory features.

Stage 3: Dimensionality Analysis

In Stage 3, a factor analysis (FA) was used to examine the correlation structure underlying certain of the features developed at Stage 2. FA has been used frequently to explore the patterns of linguistic variation detected in representative collections of texts. For example, Ervin-Tripp (as cited in Biber et al., 2004) argued that, because many important text characteristics are not well captured by individual linguistic features, investigation of such characteristics requires a focus on "constellations of co-occurring linguistic features" as opposed to individual features. FA permits easy access to such constellations by allowing patterns of linguistic co-occurrence to be analyzed in terms of underlying dimensions of variation or factors that are identified quantitatively.

In order to provide a more stable solution, the analysis was implemented with respect to entire documents, as opposed to individual paragraphs. Implementation proceeded as follows. First, candidate text features were extracted from each document in a subset of 937 texts downloaded from the GRE portion of the SourceFinder database. Each document in this subset included 1,000–5,000 words, yielding a total corpus size of more than 4.5 million words. Second, differences in text length were accounted for by re-expressing all of the count-based features as log frequency per thousand words (lfptw). Next, the major dimensions of variation underlying

the extracted features were identified by implementing a principal component analysis extraction followed by a Promax rotation. A principal component analysis extraction was selected because our primary goal involved reducing a large number of candidate features down to a more manageable number of dimension scores. A Promax rotation was selected because the resulting dimension scores were expected to be moderately correlated. Paragraph-level dimension scores were then estimated for each paragraph in the training and validation samples, and these scores were treated as additional text features in all subsequent analyses.

Stage 4: Model Development and Validation

The model development phase of the analyses was designed to generate predictions of text acceptability that closely reflected the ratings provided by the GRE test developers. Two different types of models were developed to achieve this goal: a regression model and a filtering model. The regression model was obtained by regressing paragraph acceptability (expressed on the 5-point scale) on a subset of text features, as shown in Equation 1:

$$y_i = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i \qquad (1)$$

where y_i is the average acceptability score obtained for the i[th] paragraph , the X_{ik} include the dimension scores developed at Stage 3 as well as some individual text features, the β_k are coefficients that are estimated from the available training data and the error terms, ε_i, are assumed to be independently and identically distributed with mean 0 and common variance σ^2.

Variable selection was implemented by first using a tree-based regression analysis (Brieman, Friedman, Olshen, & Stone, 1984) to select a promising subset of variables and then using a leaps analysis (Furnival & Wilson, 1974) to evaluate all possible subsets of the selected variables. Because all possible subsets of variables are evaluated, this particular estimation procedure is less subject to the variable selection problems that are symptomatic of stepwise procedures (Thompson, 1995). The resulting prediction model was then used to generate a predicted acceptability rating for each paragraph in the SourceFinder database.

One limitation of the estimation procedure described above is that it is not designed to detect violations that are serious, yet only rarely observed in the corpus. This limitation was addressed by defining a preliminary filtering step designed to detect outlying feature values, as follows. First, the training data were used to establish an acceptability range for a subset of key features. Next, any paragraph that yielded a feature value falling outside of this range was assigned a predicted acceptability score of 1 (*definitely reject*). For example, because PR passages typically vary between 90 and 130 words, the acceptability range for the paragraph length feature was specified as the range of 50–200 words, and all paragraphs with length scores falling outside of this range were assigned a predicted acceptability rating of 1 (*definitely reject*). Paragraphs that passed this preliminary filtering step were then evaluated via the regression equation described above. Thus, as this example illustrates, the preliminary filtering process was designed to capture violations that are serious enough to warrant immediate rejection, regardless of the values obtained for each of the other features.

Results

Stage 1: Data Collection

The acceptability ratings collected for the 114 paragraphs in the randomly selected portion of the training sample and the 1,000 paragraphs in the validation sample are summarized in Table 1. Table 1 illustrates the agreement between the first and second ratings obtained for each paragraph.[2] Since only a handful of ratings were obtained at Levels 2 and 4, the analysis was implemented after first collapsing Levels 1 and 2 to form a single *Reject* category and also collapsing Levels 4 and 5 to form a single *Accept* category.

The data in Table 1 suggest that the rate of exact agreement (on the collapsed scale) was fairly high in the training sample (80%) and lower in the validation sample (63%). After using the kappa statistic to adjust for agreement due to chance, as Powers (2000) recommended, these rates dropped to 70% and 45%, respectively. Variation in the experience levels of the individual raters participating at each stage of the data collection might have contributed to these differences. In particular, whereas the 4 test developers who participated in the training phase of the data collection were all highly experienced, the 14 test developers who participated in the validation phase of the data collection encompassed a much broader range of experience levels.

Table 1

Correspondence Between the First and Second Ratings Obtained for the 114 Paragraphs in the Training Sample and the 1,000 Paragraphs in the Validation Sample

	Reject (1 or 2)	Uncertain (3)	Accept (4 or 5)	Total
Training sample				
Reject (1 or 2)	84	9	1	94
Uncertain (3)	7	1	2	10
Accept (4 or 5)	0	4	6	10
Total	91	14	9	114
Validation sample				
Reject (1 or 2)	539	81	61	681
Uncertain (3)	80	31	43	154
Accept (4 or 5)	58	45	62	165
Total	677	157	166	1,000

Note. Shaded cells show the frequency of exact agreement (on the collapsed scale).

In addition to providing information about rater agreement, Table 1 also provides a sense of the overall difficulty of finding acceptable source material for use in developing new PR passages and items. The table shows that 74% of the paragraphs in the training sample were rejected by both test developers. Since the training sample was randomly selected, this suggests that a fairly large proportion of the paragraphs in the current SourceFinder database are *not* likely to be appropriate for use in developing new PR passages, and that tools designed to help test developers locate the acceptable paragraphs more quickly could lead to significant efficiency gains.

Stage 2: Feature Development

In addition to the numeric acceptability ratings described above, the raters were encouraged to provide a brief, written description of the individual text characteristics that influenced their acceptability judgments. These comments constituted the primary data considered during feature development.

Table 2 lists the individual comments collected for four paragraphs selected from the training sample. Comments provided by both raters are included. Many of the raters provided useful information about the specific aspects of text variation considered during their evaluation process. These results enabled us to develop hypotheses about the specific aspects of text variation that might prove useful for distinguishing between paragraphs scaling at the low and high ends of the acceptability scale. The comments suggested that, at a minimum, the raters tended to focus on four particular aspects of text variation: (a) level of argumentation, (b) accessibility, (c) sensitivity, and (d) content. Features designed to assess text standing on each of these four aspects of text variation are described below.

Measuring level of argumentation. The argumentation level of a paragraph is an indication of its ability to support the types of complex reasoning items needed to provide accurate measurement at the high end of the proficiency scale. Sheehan, Kostin, Futagi, Hemat, et al. (2006) noted that texts that are primarily descriptive or that merely present straight-forward exposition or narration are less likely to support challenging reasoning items, whereas texts that provide some conflict or contrast of ideas and some uncertainty about conclusions or outcomes are more likely to support such items.

Several of the comments in Table 2 address the level of argumentation aspect of text variation. In particular, the following five comments suggested that the corresponding paragraphs did not exhibit the desired level of argumentation:

1. "Not enough tension/argument."
2. "Not really any reasoning here."
3. "Not much here for our purposes (in terms of argumentation)."
4. "Not academic enough."
5. "Too thin. Descriptive rather than reasoning."

Table 2

Comments Collected for Four Training Paragraphs

Paragraph no.	Rater ID	Acceptability score	Comment[a]
3176	A	2	This would be difficult to work with because of potential sensitivity concerns and the author's fondness for jargon. [Type of violation: sensitivity, accessibility]
3176	B	1	Too much knowledge of literary theory is assumed here; also it is a pretty negative perspective on Black women writers for our purposes. [Type of violation: accessibility, sensitivity]
16949	A	1	Not enough tension/argument, and exclusive focus on religion would violate sensitivity guidelines. [Type of violation: argument, sensitivity]
16949	B	1	The discussion of religion is integral to this source, which effectively eliminates it. [Type of violation: sensitivity]
86995	C	1	Not really any reasoning here and it presumes a knowledge of a particular cultural context [Type of violation: argument, accessibility]
86995	D	2	Not much here for our purposes (in terms of argumentation) and the writing (in terms of tone, density) isn't really what we want. [Type of violation: argument]
17601	C	2	Not academic enough [Type of violation: argument, narrativity]
17601	D	1	Too thin. Descriptive rather than reasoning. [Type of violation: argument, narrativity]

[a] The aspects of text variation addressed by each comment are shown in brackets. These classifications are provided for illustrative purposes only. That is, they were not used in any of the analyses.

Features designed to measure this particular aspect of text variation were developed by considering previous research in the area of automatic genre detection. Biber (1988) defined the genre of a text in terms of its communicative purpose and argued that texts that share a common

communicative purpose tend to favor certain linguistic options over others. For example, the fact that certain verbs appear more frequently in narrative text than in expository text is believed to be a reflection of the differing communicative purposes of narrative versus expository text. According to this view, then, the fact that acceptable PR source texts share a common communicative purpose (i.e., to present a complex verbal argument in a manner that is persuasive yet still accepting of uncertainty) suggests that such texts tend to exhibit similar patterns of feature choices. Potentially useful features were determined by focusing on features that had been shown previously to be of use in distinguishing an academic rhetorical style. Examples include the frequency of abstract concept nouns (e.g., *existence* and *progress*), the frequency of verbs that appear more frequently in academic texts than in nonacademic texts (e.g., *suggest, consider*, *provide*), and the frequency of prepositions (Biber et al., 2004).

Measuring paragraph accessibility. Accessibility refers to the probability that a particular text unfairly gives advantages to an examinee with some specialized background knowledge, for example, detailed knowledge of cell biology. Several of the comments in Table 2 addressed this aspect of text variation. In particular, the following three comments suggested that the corresponding paragraphs did not exhibit the desired level of accessibility:

1. "This would be difficult to work with because of … the author's fondness for jargon."
2. "Too much knowledge of literary theory is assumed here."
3. "It presumes knowledge of a particular cultural context."

Two features that were found to be of use in measuring this particular aspect of text variation were: the number of nominalizations per 1,000 words and the type–token ratio. The number of nominalizations per 1,000 words is a count of the number of words in a candidate source paragraph that end in any of the following suffixes: *-tion*, *-ment*, or *-ity*. The analyses indicated that paragraphs with high scores on this feature were more likely to be rated as exhibiting an accessibility violation.

The type–token ratio was also observed to be of use in distinguishing paragraphs rated as exhibiting accessibility violations. This feature compares the total number of unique words in a document to the absolute number of words in the document. The first quantity (i.e., the total number of unique words in a document) is often referred to as the number of word *types*. The second quantity (i.e., the absolute number of words in a document) is often referred to as the

number of word *tokens*. Early analyses of the type–token ratio were reported in Youmans (1991). The current results suggested that texts with a high type–token ratio are more likely to be rated as exhibiting an accessibility violation.

Measuring paragraph sensitivity. A number of features designed to characterize variation relative to the sensitivity dimension of source acceptability were also evaluated. The goal of this portion of the analysis was to ensure that paragraphs rated as acceptable for use in PR passage development would not be in violation of ETS sensitivity guidelines. Two different sensitivity-related word lists were constructed. The first list is called the red flag list. It includes words and phrases that have a high probability of being present in documents rated as containing sensitivity violations and a low probability of being present in documents rated as *not* containing sensitivity violations (e.g., *abortion*, *amputated*.) The second list is called the Yellow Flag list. It contains words and phrases that are only moderately useful for detecting texts containing sensitivity violations (e.g., *addicted, depressed*).

In addition to two different sensitivity word lists, two different feature extraction approaches were considered. The first approach involved counting the number of red and yellow flag words detected in each candidate source paragraph. The second approach involved counting the number of red and yellow flag words detected in the parent articles that the candidate paragraphs were extracted from. This second approach is illustrated in Figure 1. Figure 1 shows the mean acceptability rating (average of two independent ratings) plotted versus the parent red flag word score (expressed as fptw) for the 161 paragraphs in the training sample. The small amount of vertical scatter shown at each possible acceptability score (i.e., 1.0, 1.5, 2.0, 2.5, etc.) was created by adding a small amount of random noise to each point's y-value, so that points that otherwise would be plotted on top of each other appear at distinct vertical locations. This technique is called *vertical jittering* (Chambers, Cleveland, Kleiner, & Tukey, 1983). Vertical jittering is frequently used to enhance the interpretability of two-dimensional scatter plots when the variable plotted on the vertical axis is measured on a discrete scale. The resulting plot suggests that paragraphs with high parent red flag word scores are more likely to receive ratings at the 1 or 2 level (i.e., probably or definitely reject), whereas paragraphs with low parent red flag word scores are more likely to receive ratings at the 4 or 5 level (probably or definitely accept). This suggests that the parent red flag word feature has succeeded in capturing useful information about the sensitivity status of paragraphs extracted from the SourceFinder database.

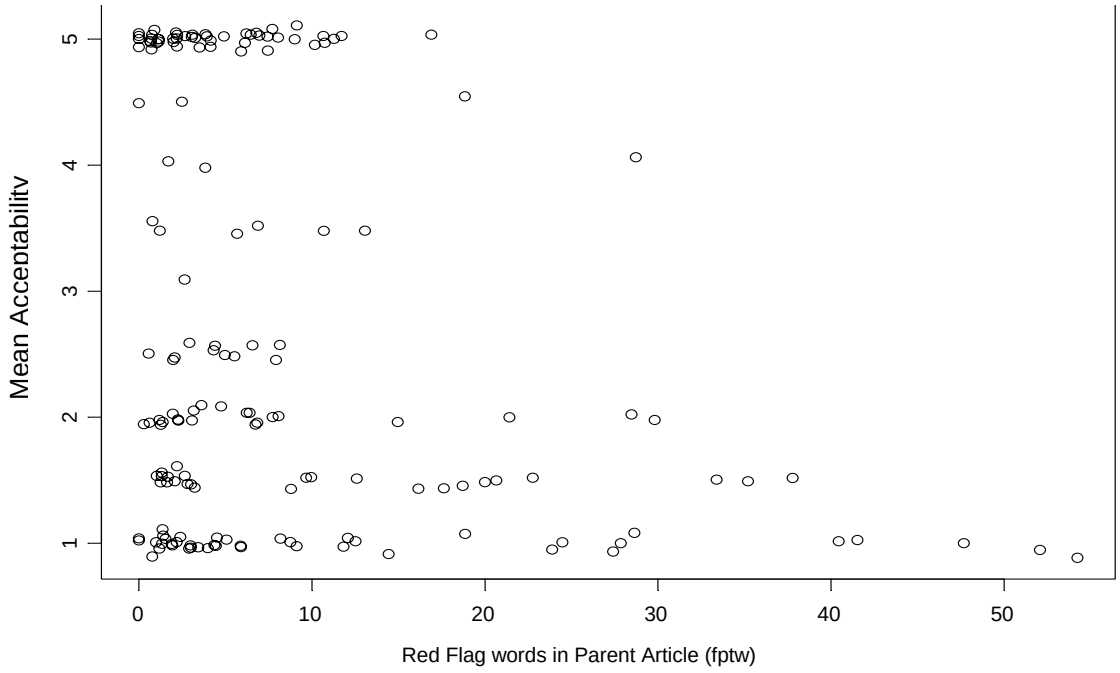


Figure 1. **Mean acceptability rating (average of two independent ratings) plotted versus parent red flag word score for the 161 paragraphs in the training sample.**

Note. n = 161 training paragraphs, fptw = frequency per thousand words.

Measuring paragraph content. Because many of the paragraphs in the SourceFinder database are relatively short (less than 200 words), content features were developed under the assumption that the content classification of a paragraph can be inherited from the content classification of the larger document from which it was extracted. Note that this assumption vastly simplifies the work needed to generate paragraph-level content classifications because it implies that the needed classifications can be developed from the existing document-level classifications.

SourceFinder's existing document-level content classifications were developed by implementing a content vector approach similar to the one described by Salton and McGill (1983) and by Salton (1989). The approach is designed to predict the content classification of a document by comparing its content words to the content words found in training passages with known content classifications. The approach is implemented in five steps, as follows:

1. Four target texts are constructed to represent the four main content areas targeted by GRE (physical science, biological science, social science, and humanities). The

target text for Content Area *k* is obtained by concatenating together previously developed passages from Content Area *k*. A total of 261 previously developed GRE passages were considered at this stage of the analysis.

2. Both the *k* target texts and the set of all source articles in the SourceFinder database are represented as vectors of standardized word frequencies. Each vector contains *w* frequency values, one for each of the *w* content words selected for consideration in the analysis. Selected content words included all of the nouns, verbs, adjectives, and adverbs detected in at least 2 of the 261 passages.

3. Because the resulting vectors were quite long, two different approaches for collapsing across rows indexed by similar content words were implemented. First, a stemming tool was used to collapse across rows associated with words arising from a common word stem. For example, *reading*, *read,* and *reads* were each represented by the single word class: *to read*. Second, a measure of word-word similarity (Lin, 1998) was used to collapse across words rated as having a high degree of distributional similarity. The similarity measure selected to implement this latter collapsing is based on Harris's (1968) distributional hypothesis, which states that words with similar meanings tend to appear in similar contexts. For example, evidence of the semantic similarity between *bacteria* and *germs* can be assembled by noting that both *bacteria* and *germs* tend to be used with the verbs *grows, lives, spreads*, and *causes* and are frequently modified by the adjectives *harmful, air borne,* and *deadly.* Based on this type of corpus evidence, *bacteria* and *germs* were rated as having a high degree of distributional similarity and were subsequently collapsed into a single word class. Note that the resulting collapsing strategy preserves part-of-speech information. That is, nouns are only collapsed with other nouns, verbs are only collapsed with other verbs, and adjectives and adverbs are only collapsed with other adjectives and adverbs.

4. The word classes defined above are then viewed as distinct dimensions of variation, and the resulting frequency vectors are viewed as observations in *t*-space, where $t < w$ is the total number of word classes remaining at the completion of the collapsing algorithm. The degree of similarity between the i^{th} parent article and the k^{th} target text is then estimated as shown in Equation 2:

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

where $S_i = [s_{i1},\ldots,s_{it}]$ is the collapsed vector of standardized term frequencies obtained for the i^{th} parent article and $G_k = [g_{k1},\ldots,g_{kt}]$ is the collapsed vector of standardized term frequencies obtained for the k^{th} target text. Here, $k = 1,\ldots, 4$ refers to the four targeted GRE content areas. This measure is called the cosine similarity measure or the cosine correlation because it is mathematically equivalent to the cosine of the angle between $S_i = [s_{i1},\ldots,s_{it}]$ and $G_k = [g_{k1},\ldots,g_{kt}]$. That is, if θ represents the angle between S_i and G_k, then $r_{ik} = \cos(\theta)$. This particular similarity measure is frequently used in text classification applications because it has been shown to be relatively insensitive to zero-frequency word classes (Leydesdorff, 2005). That is, the absence of a particular word class (e.g., *bacteria*|*germ*) does not indicate dissimilarity as strongly as the presence of a matching word class indicates similarity. This property is particularly desirable for the current application, because many of the word classes included in the *k* target vectors have only a small probability of being detected in any new document.

5. In this step, the cosine correlations obtained for the four main GRE content areas are compared, and the document is assigned to the individual content area with the maximum cosine.

To further illustrate this approach, Table 3 lists portions of the target vectors developed for the four main GRE content areas. Individual word classes found to be indicative of particular content categories are shaded. The results suggest that the GRE content areas tend to employ relatively distinct vocabularies. For example, words like *species*, *population*, *brain*, *process*, *bacteria,* and *germ* tend to occur with relatively high frequency in biological science texts and relatively low frequency in the other three types of texts. Similarly, words like *art*, *work*, *literary*, *artistic*, *writer,* and *novel* tend to occur with relatively high frequency in humanities texts and relatively low frequency in the other three types of texts.

Table 3

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

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

Note. Letters in parentheses indicate part of speech, as follows: A = adjective or adverb, N = noun, V = verb. The construction Word1|Word2 indicates words classified as having a high degree of distributional similarity. Shaded cells highlight the row-wise maximums. From *Inside SourceFinder: Predicting the Acceptability Status of Candidate Reading Comprehension Source Documents by* K. M. Sheehan, I. Kostin, Y. Futagi, R. Hemat, & D. Zuckerman, 2006 (ETS Research Report No. RR-06-24), Princeton, NJ: ETS.

The performance of this procedure relative to the task of generating content classifications for each document in the SourceFinder database was reported in Sheehan et al. (2006). That analysis demonstrated that the approach provides useful document-level content information. Because PR passages are developed from individual paragraphs, as opposed to entire documents, however, and because the "true" classification of a paragraph may not be equal to the "true" classification of its parent document, this study considered how well the existing document-level classifier performs relative to the task of generating paragraph-level content classifications.

The analysis was implemented as follows. First, a sample of 500 paragraphs that had been evaluated by the GRE test developers was obtained. Because each paragraph in this sample had been independently rated by two test developers, two content classifications were available for each paragraph. The test developers had generated these classifications after viewing the selected paragraph and, optionally, one or two of the following: the immediately preceding paragraph (if the selected paragraph had not been the initial paragraph of the document) and the immediately following paragraph (if the selected paragraph had not been the final paragraph of the document.). Thus, each human-generated classification was based on an examination of, at most, three paragraphs. Second, the two classifications obtained for each paragraph were compared, and only those paragraphs with consistent classifications were retained. This latter step yielded a reduced sample of 441 paragraphs. Next, the previously estimated document-level content classifications for these 441 paragraphs were retrieved from the SourceFinder database and were compared to the human-generated classifications. The resulting data are summarized in Table 4. Table 4 shows that even though the human-generated content classifications had been based on an examination of, at most, three paragraphs, and the corresponding automated classifications had been based on an examination of entire documents, the level of exact agreement was still relatively high. In particular, the rate of exact agreement ranged from 74% for the humanities content area to 93% for the social sciences content area. Based on these results, it was decided that the content classifications generated by the existing document-level classifier could be used to generate the paragraph-level content classifications needed for the current application. Consequently, each paragraph in the SourceFinder database was assigned the content classification developed for its parent document (i.e., the larger document that it was extracted from.)

Table 4

Automated Content Classifications Compared to the Classifications Assigned by Expert Text Developers for a Sample of 441 Paragraphs Selected From the SourceFinder Database

Classification assigned by the automated procedure	Classification assigned by the human raters				Total	% Agree
	BS	PS	SS	HU		
Biological science (BS)	52	5	10	0	67	78
Physical science (PS)	3	30	3	0	36	83
Social science (SS)	8	1	245	10	264	93
Humanities (HU)	3	1	15	55	74	74
Total	66	37	273	65	441	87

Note. Shaded cells highlight the row-wise maximums.

Figure 2 provides an additional look at a portion of these data. The plot shows 273 lines, one for each of the 273 paragraphs that had been classified consistently into the social science content area by the GRE test developers. Each line connects the four cosine measures calculated for a single paragraph. Note that almost all of the lines show a peak at the social science content area. This illustrates why the automated procedure classified 245 of the 273 "true" social science paragraphs into the "correct" content area (i.e., the content area specified by the GRE test developers).

Stage 3: Dimensionality Analysis

The preceding section documented the analyses implemented to characterize paragraph standing on four aspects of text variation: (a) level of argumentation, (b) accessibility, (c) sensitivity, and (d) content. These analyses resulted in a large number of text features, including 62 features developed to measure argumentation or accessibility, 4 features developed to measure sensitivity, and 4 features developed to measure content. Because the performance characteristics of the 62 features developed to measure argumentation or accessibility were not well understood initially, an FA was used to further investigate the aspects of text variation tapped by these features.

The FA indicated that several of the specified features were either redundant or only weakly correlated with the major dimensions of variation underlying the bulk of the features. These results were used to select a subset of 42 prime features for consideration at subsequent stages of the analysis. An FA of the 42 retained features suggested that, at most, eight dimensions of variation were being measured. The eigenvalues for these eight factors were as follows: 10.473, 4.889, 2.741, 2.132, 1.879, 1.557, 1.340, and 1.130. Since only the first six factors appeared to be construct-relevant, a six-factor solution was extracted. Taken together, these six factors accounted for nearly 60% of the shared variance.

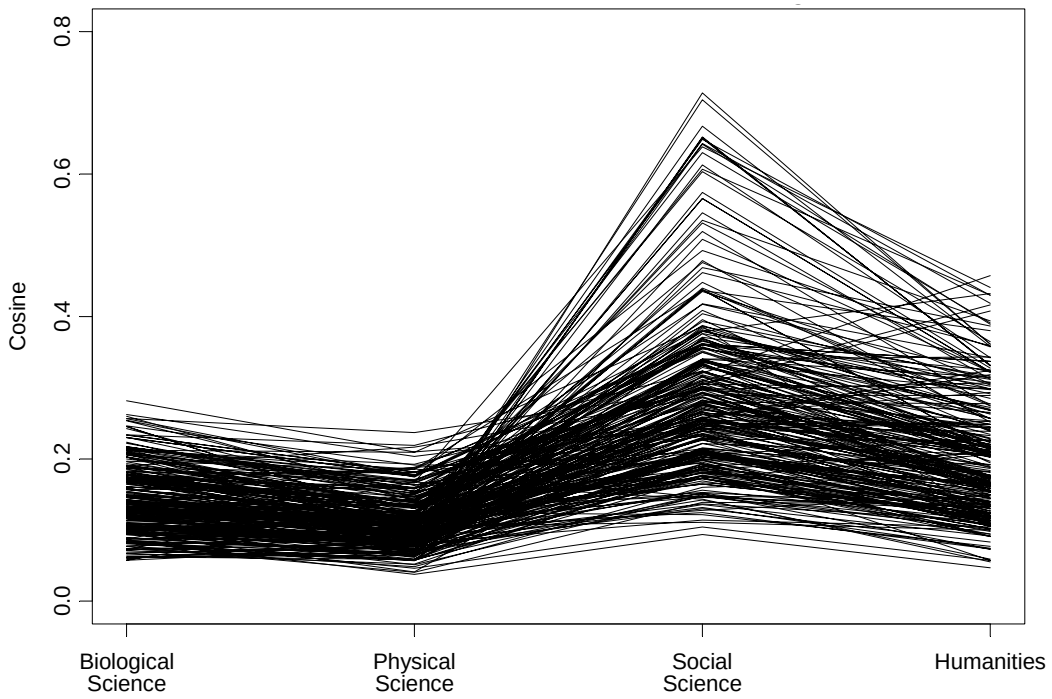


Figure 2. **Cosine similarity measures for 273 paragraphs classified by GRE test developers as belonging to the social science (SS) content area.**

To further illustrate these results, Table 5 lists the individual features that were most strongly associated with each factor. The factor loadings obtained in the Promax rotation are also shown. Loadings near +1.0 are indicative of strong positive feature–dimension associations; loadings near -1.0 are indicative of strong negative feature–dimension associations.

Table 5

Dimensions of Variation Detected in Candidate GRE Source Texts With Associated Linguistic Features and Loadings

Dimension/feature	Loading
Dimension 1: Narrative discourse	
Words enclosed in quotes (lfptw)	+.86
Communication verbs (lfptw) [e.g., *say, shout, call*]	+.84
Contractions (lfptw)	+.82
Second person pronouns (lfptw)	+.79
First person singular pronouns (lfptw) [e.g., *I, me, mine*]	+.75
Mental state verbs (lfptw) [e.g., *appreciate, enjoy*]	+.72
Indefinite pronouns (lfptw) [e.g., *anybody, everybody*]	+.61
Third person singular pronouns (lfptw) [e.g., *he, she, his*]	+.55
Fiction verbs (lfptw) [e.g., *sit, walk, look, ask, tell*]	+.53
Dimension 2: Academic versus nonacademic discourse	
Cognitive process/perception nouns (lfptw) [e.g., *concept*]	+.96
Research words (lfptw) [e.g., *experiment, hypothesis*]	+.92
Abstract concept nouns (lfptw) [e.g., *existence, progress*]	+.75
Nominalizations (lfptw) [words ending in *-tion, -ment, -ness,* and *-ity*]	+.74
Average word length (log characters)	+.69
Causal reasoning words (lfptw) [e.g., *imply, implicate*]	+.59
Academic conjuncts (lfptw) [e.g., *conversely, furthermore*]	+.50
Academic verbs (lfptw) [e.g., *suggest, consider, provide*]	+.45
Time adverbs (lfptw) [e.g., *soon, yesterday, today, once*]	-.65
Dimension 3: Overt expression of argumentation	
Possibility modals (lfptw) [e.g., *can, can't, could, couldn't*]	+.77
Prediction modals (lfptw) [*shall, will, won't, wouldn't*]	+.61
Other modals (lfptw) [*possibly, certainly*]	+.58
Conditional subordinators (lfptw) [*if, unless*]	+.57
Belief adverbs (lfptw) [*clearly, fortunately*]	+.52
To infinitives (lfptw)	+.42

(Table continues)

Table 5 (continues)

Dimension/feature	Loading
Dimension 4: Oppositional reasoning	
Analytic negation (lfptw) [*not*]	+.67
Academic downtoners (lfptw) [e.g., *hardly, barely, merely*]	+.65
Oppositional reasoning words (lfptw) [e.g., *argue, challenge*]	+.54
Synthetic negation (lfptw) [*no*]	+.52
Nonacademic negations (lfptw) [e.g., *never, none, nothing*]	+.49
Nouns (lfptw)	-.45
Dimension 5: Sentence complexity	
Median of longest chunk (log words in longest chunk)	+.88
Average sentence length (log words)	+.87
Adjectives (lfptw)	+.51
Words not in EWFG (lfptw)	-.61
Dimension 6: Unfamiliar cocabulary	
No. of unique words (not including proper nouns) not in EWFG (lfptw)	+.78
Type–token ratio (estimated from the first 400 words only)	+.53
Average EWFG word frequency value	-.63

Note. lfptw = log frequency per thousand words; EWFG = *Educator's Word Frequency Guide*, a word frequency index published by Touchstone Applied Science Associates (Zeno, Ivens, Millard, & Duvvuri, 1995).

Table 5 also provides a short descriptive label for each dimension. These were developed by considering the pattern of variation implied by the highly weighted features within each dimension. (Only features with loadings in excess of .35, a total of 42 features, were considered during the process of developing dimension labels.) The results suggest that the argumentation and accessibility levels of a text can be decomposed into six underlying dimensions of variation: (a) the degree of narrative orientation detected in a text (narrative discourse), (b) the extent to which the text exhibits features that are more characteristic of academic discourse than of nonacademic discourse (academic versus nonacademic discourse), (c) the amount of overt argumentation detected (overt expression of argumentation), (d) the amount of opposition

detected (oppositional reasoning), (e) the complexity level of a text's component sentences (sentence complexity), and (f) the vocabulary level of the text (unfamiliar vocabulary).

To better understand the aspects of text variation captured by the resulting dimension scores, Figure 3 shows how the third and fourth dimensions (i.e., overt expression of argumentation and oppositional reasoning) relate to the paragraph acceptability ratings provided by the GRE test developers. The top panel shows the mean acceptability rating (average of two independent test developer ratings) plotted versus the overt expression of argumentation score for the 161 paragraphs in the training sample (including both the 114 randomly selected paragraphs and the 47 historical paragraphs.) The bottom panel shows a similar display for the oppositional reasoning score. In each plot, vertical jittering is used to improve interpretability (Chambers et al., 1983). The resulting plots suggest that both the overt expression of argumentation factor and the oppositional reasoning factor are useful for distinguishing among paragraphs with low versus high mean acceptability ratings. In particular, paragraphs with high argument or oppositional reasoning scores are more likely to receive acceptability ratings of 4 or 5 (i.e., probably accept or definitely accept), whereas paragraphs with low argument or oppositional reasoning scores are more likely to receive lower acceptability ratings (i.e., probably reject or definitely reject). The relationship is not that precise, however, suggesting that other factors in addition to those considered in the current plots are also likely to be important.

Stage 4: Model Development and Validation

Two models were developed: (a) a preliminary filtering model, and (b) a regression model designed to predict paragraph acceptability conditional on the available text features. These models were used to generate a predicted paragraph acceptability score, expressed on the 1–5 scale, for each paragraph in the validation sample. Two different approaches for validating the resulting acceptability predictions were then implemented. The first approach considered the agreement between the reject, uncertain, and accept classifications provided by the GRE test developers (after collapsing) and the corresponding classifications generated via the predicted acceptability ratings. The second approach considered differences in the operating characteristic (OC) curves estimated for a paragraph development task implemented with and without access to the predicted paragraph acceptability ratings. Results obtained via each approach are summarized below.

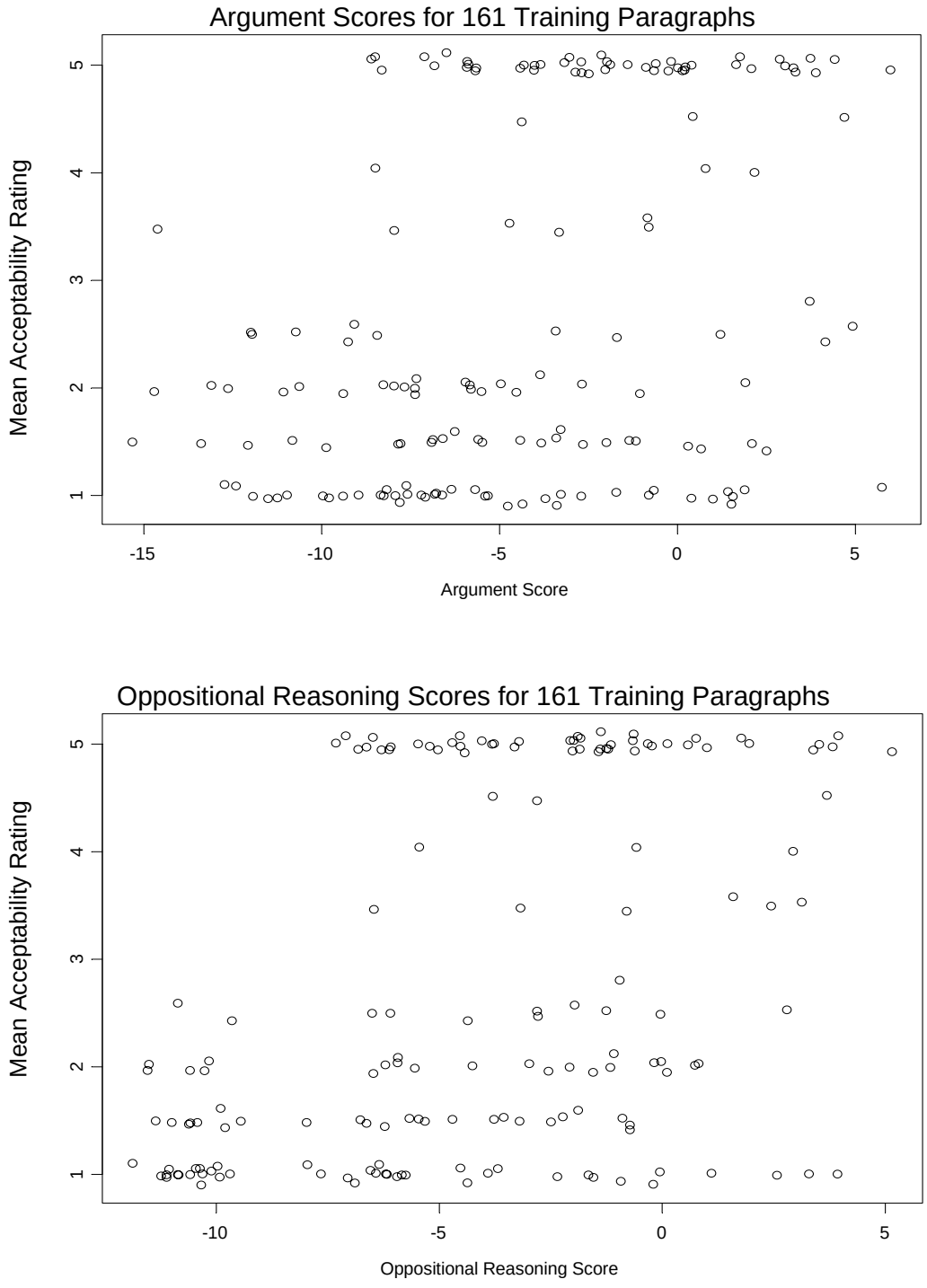


Figure 3. **Mean acceptability rating plotted versus Overt Expression of Argumentation score (top panel) and Oppositional Reasoning score (bottom panel) for the 161 paragraphs in the training sample.**

Validation Approach 1: Calculating the agreement between human and automated classifications of paragraph acceptability. One difference between the paragraph acceptability ratings provided by the GRE test developers and those generated via the automated procedure described above is that the former are expressed on an ordinal scale (i.e., $\hat{y}_{TD} \in$ {Reject, Uncertain, Accept}), whereas the latter are expressed on a continuous scale (i.e., $1 \leq \hat{y}_{SE} \leq 5$). To permit analyses of the level of agreement between these two sets of results, a transformation function was developed to reexpress the continuously scaled automated ratings on the test-developer-preferred reject-uncertain-accept scale. To preserve the independence of the validation sample, the required transformation function was developed from the training data only. The expected percentage of acceptable documents was also preserved by restricting the sample to only those training documents that had been randomly selected from the SourceFinder database.

The required transformation function was then obtained as follows. First, the available randomly-selected training paragraphs (n = 114) were divided into three groups based on the average of their two test-developer ratings: (a) Paragraphs with an average rating less than or equal to 2 were classified into the Reject category, (b) paragraphs with an average rating greater than or equal to 4 were classified into the Accept category, and (c) the remaining paragraphs were classified into the Uncertain category. This process yielded a total of 89 rejected paragraphs, 19 uncertain paragraphs, and 6 accepted paragraphs. Next, the training paragraphs were ordered from least acceptable to most acceptable based on the continuously scaled ratings yielded by the automated procedure. Finally, two cut points were selected: one to distinguish between a predicted rating of reject and a predicted rating of uncertain, and one to distinguish between a predicted rating of uncertain and a predicted rating of accept. These cut points were selected so that the resulting marginal distribution reflected that obtained from the test developer ratings. That is, the reject–uncertain cut point was selected so that the automated procedure would reject exactly 89 paragraphs, and the uncertain–accept cut point was selected so that the automated procedure would accept exactly six paragraphs. The numerical cut points needed to obtain these results were 3.13 and 4.05, respectively. The required transformation function was thus specified as follows:

$$\text{If } \hat{y}_{SE} \leq 3.13 \quad \text{then SourceFinder class = reject}$$

$$\text{If } \hat{y}_{SE} \geq 4.05 \quad \text{then SourceFinder class = accept}$$

Otherwise, SourceFinder class = uncertain

Note that this approach is designed to maximize the level of exact agreement between the human and automated ratings in the training sample. Table 6 summarizes the resulting agreement information

Table 6

Agreement Between Human and Automated Paragraph Acceptability Classifications for 114 Training Paragraphs

Mean test developer rating	Automated acceptability rating			Total
	Reject ≤ 3.13	Uncertain 3.13–4.05	Accept ≥ 4.05	
Reject, 1–2	77	11	1	89
Uncertain, 2.5–3.5	11	5	3	19
Accept, 4+	1	3	2	6
Total	89	19	6	114

Note. Shaded cells highlight the frequency of exact agreement.

The procedure outlined above was then used to generate an automated acceptability classification expressed on the collapsed scale (i.e., accept, uncertain, or reject) for each of the 1,000 paragraphs in the validation sample. Table 7 shows how the resulting ratings compared to the first set of human ratings. The table confirms that the ratings generated by the automated procedure agreed with those provided by the human raters 62% of the time when human performance was characterized via the first set of human ratings. A similar table was constructed to characterize performance relative to the second set of human ratings. This second table indicated that that the ratings generated by the automated procedure agreed with those provided by the human raters 61% of the time when human performance was characterized via the second set of human ratings. These agreement rates compare favorably to the human-to-human agreement rate previously shown in Table 1. That is, for the 1,000 paragraphs in the validation sample, the level of exact agreement between two human raters was 63%, and that between SourceFinder and a human rater was 61–62%. Thus, these results suggest that, when the

collapsed scale is used, the level of exact agreement between SourceFinder and a human rater is nearly indistinguishable from that between two human raters.

Table 7

Agreement Between Human and Automated Paragraph Acceptability Classifications for 1,000 Validation Paragraphs

First test developer rating	Automated acceptability rating			Total
	Reject ≤ 3.13	Uncertain 3.13 – 4.05	Accept ≥ 4.05	
Reject, 1- 2	514	83	84	681
Uncertain, 2.5 – 3.5	76	31	47	154
Accept, 4 - 5	64	23	78	165
Total	654	137	209	1,000

Note. Shaded cells highlight the frequency of exact agreement.

The results in Table 7 can also be used to compare the percent of acceptable paragraphs located with and without the aid of the automated acceptability predictions generated in this study. For the "Without Automated Filtering" condition, the acceptance rate is calculated as follows:

Acceptance Rate Without Filtering = 165/1000 = 16.5%

To obtain a corresponding set of results for the "With Automated Filtering" condition, we began by restricting the sample to just those paragraphs that the automated tool classified at the accept level. The acceptance rate in this subsample was then calculated as follows:

Acceptance Rate With Filtering = 78/209 = 37.3%

These results suggest that the test developers can increase the proportion of acceptable paragraphs located per unit time interval by confining their attention to only those paragraphs that the new SourceFinder module classifies at the accept level. Note, however, that this approach also yielded 87 false negatives. That is, the approach screened out 87 paragraphs that, based on the human ratings, are actually acceptable. The proportion of truly acceptable

documents that are correctly classified as acceptable by an automated text analysis procedure is frequently referred to as its *recall rate* (Van Rijsbergen, 1979). The data in Table 7 suggest that SourceFinder's paragraph-level recall rate is 78/165, or 47%. Note that this relatively low recall date is due, in part, to the relatively low level of human to human agreement

Validation Approach 2: Comparison of OC curves. The new SourceFinder module was also evaluated by comparing two different OC curves: (a) an OC curve developed to characterize relevant performance characteristics of the automated rating procedure and (b) a baseline OC curve designed to characterize performance when automated source acceptability ratings are *not* considered. In each case, model performance is characterized by plotting the probability of finding an acceptable source paragraph conditional on a particular approach for sorting the available candidates. For the automated procedure, candidate paragraphs are sorted from most acceptable to least acceptable, as determined from the continuously scaled acceptability ratings provided by the estimated acceptability model. For the baseline model, a random sort order is used. The rational for this approach is that, since candidate source paragraphs are presented to SourceFinder users as a sorted list, the greatest reductions in source evaluation times will be achieved when all of the truly acceptable paragraphs are sorted to the beginning of the list and all of the truly unacceptable paragraphs are sorted to the end of the list. The OC curves calculated as described above provide a graphical view of the expected proportion of acceptable documents located for paragraphs situated at any point in the sort list (e.g., near the beginning of the list or near the end of the list). A desirable OC curve is high for documents near the beginning of the list and low for documents near the end of the list

This approach was implemented as follows. First, the automated procedure was used to generate a continuously scaled acceptability score for each paragraph in the validation sample. Second, these scores were used to order all of the paragraphs in the validation sample from high to low, that is, from those rated as most acceptable for use in developing a new PR stimulus paragraph to those rated as least acceptable for that task. Next, the sorted paragraphs were divided into groups such that the first group contained the 50 paragraphs with the highest acceptability scores, the second group contained the 50 paragraphs with the next highest acceptability scores, and so on. Since there were a total of 1,000 paragraphs in the validation sample, this process yielded a total of 20 groups. The proportion of acceptable-rated paragraphs in each group was then determined, and these proportions were plotted versus group rank order.

For this analysis, a paragraph was considered to be acceptable only if the average of its two human ratings were at least 4.

The resulting OC curves are shown in Figure 4. For purposes of comparison, the plot also shows the OC curve obtained when a random process is used to establish the assignment of paragraphs to groups. Differences in the resulting OC curves suggest that, relative to a random process, the automated procedure sorts a higher proportion of the truly acceptable paragraphs to the beginning of the list and a much lower proportion of the truly acceptable documents (close to 0%) to the end of the list.

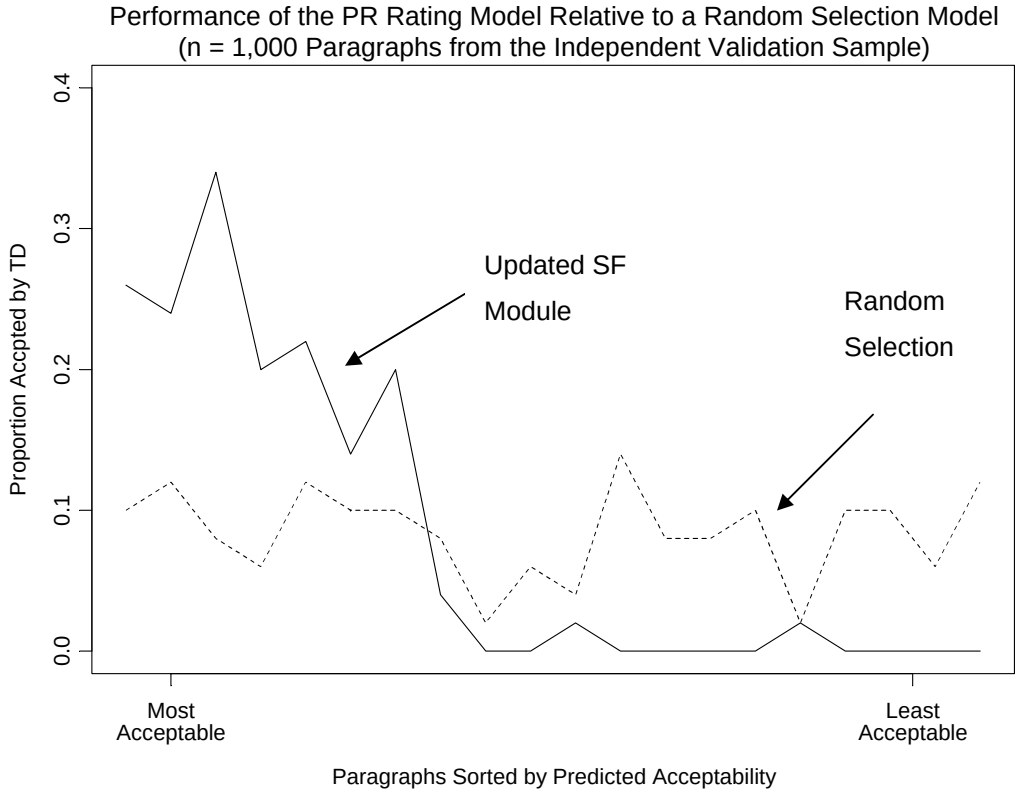


Figure 4. **Comparison of two different operating characteristic curves, one for the updated SourceFinder module and one for a random selection model.**

Note. PR = paragraph reading; SF = SourceFinder; TD = test developer.

The OC curves in Figure 4 also may be used to compare the acceptance rates expected under the estimated acceptability model to those expected under a random sort model. The plot shows that a strategy of viewing paragraphs randomly is likely to yield an acceptance rate of

about 10%, whereas a strategy of viewing only those paragraphs that received high ratings from the updated SourceFinder module yields an acceptance rate closer to 30%. This increase should translate directly into efficiency gains. That is, the strategy of presenting candidate paragraphs to SourceFinder users via the sort order determined from the updated automated acceptability module means that fewer unacceptable sources will have to be evaluated in order to find the desired number of acceptable sources.

Conclusions

A major difference between the goals of this study and those considered in previous SourceFinder research is that, although this study was designed to provide paragraph-level estimates of source acceptability, all of the previous research in this area has focused on providing document-level estimates of source acceptability. One consequence of this difference is that, whereas the source texts considered in the previous research were all relatively lengthy, the source texts considered in this study were all relatively short. The analyses confirmed, however, that even for these shorter texts, test developers can save a considerable amount of search time by confining their attention to only those documents that SourceFinder rated as having a relatively high probability of being acceptable for use in developing a new PR stimulus passage. In particular, the analyses confirmed that it is possible to increase the percent of acceptable documents located per unit time interval from the current rate of about 10% to nearly 30%, an increase that should translate directly into significant efficiency gains.

The analyses also considered the level of exact agreement between the paragraph acceptability ratings generated via the updated SourceFinder module and those provided by the human raters. These analyses confirmed that, when paragraph acceptability is expressed on a collapsed scale (i.e., reject, uncertain, and accept), the level of exact agreement between SourceFinder and a human rater is nearly indistinguishable from that between two human raters. In particular, the level of exact agreement between SourceFinder and a human rater ranged from 61% to 62%, whereas the level of exact agreement between two human raters was only slightly higher at 63%.

This study also provided useful information about the aspects of text variation considered by GRE test developers during the process of searching for acceptable stimulus materials. In particular, the analyses confirmed that the following four aspects of text variation are frequently

considered during the process of evaluating candidate source paragraphs: (a) level of argumentation, (b) accessibility, (c) sensitivity, and (d) content.

A novel approach for measuring the level-of-argumentation aspect of text variation was also introduced. The approach involved decomposing level of argumentation into six separable dimensions or factors of variation, as follows: (a) Narrative Discourse, or the degree of narrative orientation detected in a text; (b) Academic Versus Nonacademic Discourse, or the extent to which the text exhibits features that are more characteristic of academic discourse than of nonacademic discourse; (c) Overt Expression of Argumentation, or the amount of overt argumentation detected; (d) Oppositional Reasoning, or the amount of opposition detected; (e) Sentence Complexity, the complexity level of a text's component sentences and (f) Unfamiliar Vocabulary, the vocabulary level of the text. An automated approach for assessing each component was then developed.

An automated approach for measuring the content dimension of source acceptability was also developed. This approach employed a content vector analysis (Salton, 1989; Salton & McGill, 1983) to compare the content words found in each new document to the content words found in a set of training passages with known content classifications. The analysis confirmed that the approach yields content classifications that exhibit acceptable levels of agreement with the content classifications provided by GRE test developers.

Study 2: Facilitating More Effective Item Targeting

Item development efficiency also may be enhanced by helping item writers to generate new items that are optimally configured to provide unambiguous evidence about examinees' mastery status on targeted combinations of skills; these items, as a result, would be more likely to scale at targeted difficulty levels. This section describes the analyses implemented to support that goal.

Method

Evidence-Centered Design

Our analysis of alternative item targeting techniques was informed by the test design framework described by Mislevy, Steinberg, and Almond (2002) and by Mislevy, Almond, and Lukas (2003). Called evidence-centered design (ECD), this framework highlights the link between manipulable task features and the knowledge, skills, and abilities that are the true targets

of inference. The approach involves documenting relationships among observable task features, hypothesized processing requirements, and subsequent variation in item difficulty by specifying three structures: (a) a student model, (b) a task model, and (c) an evidence model. Mislevy et al. (2003) defined these structures as follows.

The student model. Every assessment is designed to support inferences about some subset of examinee proficiencies. In the ECD framework, critical student-level variables are stored in the student model. At any given point in time, these variables characterize the assessor's current beliefs about examinees' mastery status on each of the knowledge, skills, and abilities that bear on the intended inferences.

The task model. Every item taps some array of skills. In the ECD framework, critical task-level variables are stored in the task model. These variables characterize the features of items that are expected to be indicative of differences in required skills.

The evidence model. This model describes how beliefs about student model variables should be updated as tasks with known properties are administered and scored.

The ECD framework, as described above, was originally developed for use in designing *new* assessments. To that end, the framework encourages test designers to begin their work by first specifying the claims about examinees' proficiencies that the new test will be designed to support, then specifying the types of observed evidence necessary in order to support such claims, and finally specifying the observable characteristics of tasks appropriate for use in eliciting the desired evidence. Thus, when the goal of an ECD analysis is to develop a blueprint for a new assessment, researchers are encouraged to specify first a student model; then an evidence model; and finally, at the culmination of the design process, one or more task models.

The goals of the current study were somewhat different: Instead of designing a *new* assessment, we were interested in investigating the skills needed to score at increasingly advanced levels on an *existing* assessment. The methodology developed to achieve this goal is summarized below. Three stages are identified: (a) data collection, (b) hypothesis generation about required skills, and (c) model development and validation.

Stage 1: Data Collection

Because the PR item type is a new item type, the data collection effort succeeded in gathering a total of just 125 items. This set represented all of the PR items that had been written, pretested, and reviewed as of January 2005. Item pretesting procedures were designed to ensure

that (a) at least 900 examinees responded to each item, (b) all of the participating examinees belonged to the targeted population, and (c) all of the participating examinees were sufficiently motivated.

The data collected for each item included two statistics that are routinely considered in item pretesting studies: the item delta statistic and the item-test biserial correlation. The item delta statistic, δ_i, is obtained by transforming the item proportion correct statistic, p_i, as shown in Equation 3:

$$\delta_i = 4\ \Phi^{-1}\ (1 - p_i\ ) + 13 \qquad (3)$$

where Φ^{-1} is the inverse normal transformation that transforms a probability value into a normal deviate with unit variance. The delta for an item is thus "that value on the baseline of a normal curve with mean 13.0 and standard deviation 4, above which the area under the curve is equal to the proportion passing the item" (Donlon & Angoff, 1971, p. 23).The resulting statistic has the following desirable properties: Larger values indicate more difficulty, and smaller values indicate less difficulty. Thus, item stimulus features that contribute to increases in item difficulty should yield positive correlations with item delta, and item stimulus features that contribute to decreases in item difficulty should yield negative correlations with item delta.

Item-test biserial correlations were also obtained for each item. This statistic is calculated as the correlation between the dichotomous (0/1) score obtained on the i^{th} item and the total score obtained on the entire verbal section. Test developers routinely use this measure to screen out items that have unusually low correlations with the proficiencies measured by the test as a whole.

Two types of potentially relevant item classifications were also collected. Item type classifications were inference (IN), primary purpose (PP), rhetorical purpose (RP), or vocabulary in context (VC). Item format classifications were multiple choice (MC), highlight sentence (HS), or select all correct options (SA). Sample items selected to illustrate these classifications are shown in Figures 5 and 6.[3] The item type and format classifications assigned to these five sample items are shown in Table 8. Table 8 also identifies the correct option choice for each item (also called the key.) Although only one correct option is listed for Item 1, the sample SA item, items written in this format could have one, two, or three correct options, and full credit is only awarded when the truth status of all three options has been indicated correctly.

Table 8

Item Type and Format Classifications for Five Sample Items

Passage	Item no.	Item type[a]	Response format[b]	Correct response
A	1	IN	SA	Answer (c)
A	2	VC	MC	Answer (a)
B	3	PP	MC	Answer (e)
B	4	IN	MC	Answer (c)
B	5	RP	HS	Sentence 5

[a.]IN = inference; PP = primary purpose; RP = rhetorical purpose; VC = vocabulary in context.[b]HS = highlight sentence; MC = multiple choice; SA = select all correct.

Passage A:

Scholarship on political newspapers and their editors is dominated by the view that as the United States grew, the increasing influence of the press led, ultimately, to the neutral reporting from which we benefit today. Pasley considers this view oversimplified, because neutrality was not a goal of early national newspaper editing, even when editors disingenuously stated that they aimed to tell all sides of a story. Rather, the intensely partisan ideologies represented in newspapers of the early republic led to a clear demarcation between traditional and republican values. The editors responsible for the papers' content — especially those with republican agendas — began to see themselves as central figures in the development of political consciousness in the United States.

1. *Consider each of the choices separately and select all that apply.*
   The passage suggests that Pasley would agree with which of the following statements about the political role of newspapers?

   A. Newspapers today are in many cases much less neutral in their political reporting than is commonly held by scholars.
   B. Newspapers in the early United States normally declared quite openly their refusal to tell all sides of most political stories.
   C. The editorial policies of some early United States newspapers became a counterweight to proponents of traditional values.

2. In the context in which it appears, "disingenuously" most nearly means

   A. insincerely
   B. guilelessly
   C. obliquely
   D. resolutely
   E. pertinaciously

***Figure 5.* Sample items selected to illustrate the item formats of select all correct (Item 1) and vocabulary in context (Item 2).**

Passage B

In *Raisin in the Sun,* Lorraine Hansberry does not reject integration or the economic and moral promise of the American dream; rather, she remains loyal to this dream while looking, realistically, at its incomplete realization. Once we recognize this dual vision, we can accept the play's ironic nuances as deliberate social commentaries by Hansberry rather than as the "unintentional" irony that Bigsby attributes to the work. Indeed, a curiously persistent refusal to credit Hansberry with a capacity for intentional irony has led some critics to interpret the play's thematic conflicts as mere confusion, contradiction, or eclecticism. Isaacs, for example, cannot easily reconcile Hansberry's intense concern for her race with her ideal of human reconciliation. But the play's complex view of Black self-esteem and human solidarity as compatible is no more "contradictory" than Du Bois's famous, well-considered ideal of ethnic self-awareness coexisting with human unity, or Fanon's emphasis on an ideal internationalism that also accommodates national identities and roles.

3. The author's primary purpose in the passage is to

   A. explain some critics' refusal to consider *Raisin in the Sun* a deliberately ironic play
   B. suggest that ironic nuances ally *Raisin in the Sun* with Du Bois's and Fanon's writings
   C. analyze the fundamental dramatic conflicts in *Raisin in the Sun*
   D. emphasize the inclusion of contradictory elements in *Raisin in the Sun*
   E. affirm the thematic coherence underlying *Raisin in the Sun*

4. The author of the passage would probably consider which of the following judgments to be most similar to the reasoning of the critics described in the highlighted sentence?

   A. The world is certainly flat; therefore, the person proposing to sail around it is unquestionably foolhardy.

   B. Radioactivity cannot be directly perceived; therefore, a scientist could not possibly control it in a laboratory.

   C. The painter of this picture could not intend it to be funny; therefore, its humor must result from a lack of skill.

   D. Traditional social mores are beneficial to culture; therefore, anyone who deviates from them acts destructively.

   E. Filmmakers who produce documentaries deal exclusively with facts; therefore, a filmmaker who reinterprets particular events is misleading us.

5. Click on the sentence in the passage in which the author provides examples that reinforce an argument against a critical response cited earlier in the passage.

Figure 6. **Sample items selected to illustrate the primary purpose, inference, and rhetorical item types.**

Stage 2: Feature Development

Kintsch's (1988, 1998) construction-integration model of reading comprehension provided a starting point for the specification of critical task features and associated required skills. This model characterizes comprehension in terms of two reader-developed text representations: the *textbase* and the *situation model*. The textbase is a representation of the text that preserves the meaning, but not the exact wording or syntax, of the propositions presented in a text. The situation model is an extended representation of the text that is created by integrating the textbase with relevant prior knowledge and experiences. Kintsch (1988) noted that the type of processing engaged in during the creation of a situation model for a particular text will vary according to the genre of the text and the reader's goals.

The construction-integration model suggests that breakdowns in comprehension are most likely to occur when a reader is unable to create an accurate textbase or is unable or not motivated to create a rich situation model. Since GRE examinees are likely to be highly motivated, this study focuses on text characteristics that may either facilitate or impede an examinee's ability to develop a mental representation that is sufficiently rich to distinguish among the various options presented with an item. The identified item stimulus features and associated skills are then evaluated by implementing the model development and validation approach described below.

Stage 3: Model Development and Validation

In Stage 3, a tree-based regression approach (Brieman et al., 1984) is used to validate the hypotheses about critical task features and associated skills that were developed at Stage 2. Researchers frequently have noted the crucial role of item difficulty modeling for validating hypotheses about required skills (Bejar, 1993; Embretson, 1998; Embretson & Gorin, 2001; Huff, 2006; Sheehan, 1997; Sheehan, Kostin, & Persky, 2006; Sheehan & Mislevy, 1990). Tree-based regression models are frequently used for this purpose because they are particularly effective at accounting for nonlinear item attributes (Enright & Sheehan, 2002; Sheehan, 1997; Sheehan, Kostin, Futagi, Hemat, et al., 2006).

Two different tree-based regression models were estimated: one to evaluate the skill classifications developed for the VC items, and one to evaluate the skill classifications developed for the IN, PP, and RP items. The item delta statistic served as the dependent variable considered in each model. Tree models are fit in a forward stepwise fashion. That is, the analysis begins

with all items classified as measuring a single undifferentiated skill. Potential improvements to this model are then evaluated by using a recursive partitioning algorithm (Brieman et al., 1984) to estimate the reductions in unexplained difficulty variation resulting from all possible splits of all possible predictor variables. At each stage of the analysis, the effects of the available predictor variables are evaluated using deviance, a statistical measure of the unexplained variation remaining after each new variable is added to the model. In the analyses reported in this paper, deviance is calculated as the sum of squared differences between the observed and predicted values of the item delta statistic.

Tree-based model-fitting algorithms are designed to continue splitting the available observations until a predefined stopping rule has been triggered. The stopping rule defined for the current analyses specified that model fitting would continue until the node size was reduced to 10 or fewer items or until additional reductions in deviance were not possible. Once the stopping rule is triggered, the model-fitting process ends, and the mean value of item difficulty within each terminal node is taken as the predicted value of item difficulty for each of the items in each of the nodes. The more homogeneous the node is, the more accurate the prediction will be. Thus, the terminal node definitions resulting from a tree-based regression analysis provide an item clustering scheme that minimizes within-cluster variation while simultaneously maximizing between-cluster variation. This clustering scheme also indicates the task features appropriate for use in developing a student model, a task model, and an evidence model. Thus, two types of results are provided: (a) detailed information about the skills needed to respond correctly to increasingly more difficult items and (b) a blueprint for developing new items that are optimally constructed to provide high-quality evidence about targeted proficiencies.

Results

Stage 1: Data Collection

All of the item pretest data considered in this paper were collected in a series of pilot tests administered in 2004. Resulting item analysis statistics are summarized in Table 9, both for the full item set and for items classified by item format and type. The item deltas ranged from a low of 3.7 to a high of 18.6, and the item-test biserial correlations ranged from a low of -0.17 to a high of 0.65.

Table 9

Item Analysis Statistics for 125 PR Items Classified by Item Format and Type

	#	Item delta			Item-test biserial correlation					Total[c] items
		Min.	Mean	Max.	Min.	Mean	Max.	No. < .2	% < .2	
Response format[a]										
HS	6	3.7	9.6	12.8	.26	.42	.64	0	0	6
MC	90	6.0	12.6	17.4	-.17	.35	.65	13	14	77
SA	29	10.1	14.9	18.6	.01	.26	.53	8	28	21
Item type[b]										
VC	24	6.0	11.6	17.3	.14	.42	.65	1	4	23
RP	16	9.3	11.8	15.6	.19	.37	.58	1	6	15
IN/PP	85	3.7	13.6	18.6	-.17	.30	.64	19	22	66
Response format = MC (n = 90)										
VC	23	6.0	11.3	17.3	.14	.42	.65	1	4	22
RP	13	9.3	11.4	13.4	.19	.39	.58	1	8	12
IN/PP	54	9.0	13.4	17.4	-.17	.31	.54	11	20	43
Response format = SA (n = 29)										
VC	1	16.7	16.7	16.7	.53	.53	.53	0	0	1
RP	1	15.6	15.6	15.6	.28	.28	.28	0	0	1
IN/PP	27	10.2	14.8	18.6	.01	.24	.49	8	30	19
All items	125	3.7	13.0	18.6	-.17	.33	.65	21	17	104

[a] HS = highlight sentence; MC = multiple choice; SA = select all correct. [b] VC = vocabulary in context, RP = rhetorical purpose, IN = inference; PP = primary purpose. Because the dataset included only 4 PP items, the PP category has been collapsed with the IN category. This collapsing strategy reflects the belief that PP items are likely to be more like IN items than like RP items because, while RP items involve inferences about text structure, both PP items and IN items involve inferences about text content. [c]The total number of items in each category that were included in the difficulty analyses.

Table 9 also provides a count of the number of items with biserial correlations below 0.20, a cut-off value that is routinely used to filter out unsatisfactory items. Since items with biserial correlations below 0.20 are not used operationally, they were also excluded from the model development activities reported below. Thus, the final dataset included 104 items; of these, 77 were written in the traditional MC format, 21 were written in the new SA format, and 6 were written in the new HS format.

Because the VC items appeared to be tapping a unique set of skills, Stages 2 and 3 were implemented twice: once for the VC items and once for the IN, PP, and RP items. Resulting information about critical task features and associated required skills is summarized below.

Stage 2: Feature Development for VC Items

Our analysis of the VC items suggested that examinees' tendencies toward correct response could be explained by considering mastery status on the following proficiencies: (a) the ability to retrieve needed vocabulary knowledge from long-term memory and (b) the ability to understand a common word used in an uncommon sense. These two proficiencies are associated with two very different evidentiary paradigms: capturing evidence about *past reasoning*, and capturing evidence about *current reasoning*. In particular, since vocabulary knowledge tends to develop over time, items designed to collect evidence about students' mastery status relative to the first proficiency are most accurately characterized as belonging to the first evidentiary paradigm, capturing evidence about cognitive operations implemented in past interactions with text. By contrast, since comprehension of a common word used in a novel sense involves current reasoning, items designed to tap this second proficiency are more properly characterized as belonging to the second evidentiary paradigm, capturing evidence about cognitive operations implemented at the moment of testing.

The approach developed to characterize item demands relative to the first proficiency, retrieving needed vocabulary knowledge from long-term memory, was based on information extracted from the *Educator's Word Frequency Guide* (EWFG), a compilation of word frequency information developed by Touchstone Applied Science Associates (Zeno et al., 1995). This index was developed from a corpus of 17 million words distributed across 60,527 text samples selected from textbooks, works of literature, and popular works of fiction and nonfiction used in schools and colleges throughout the United States. At the time that it was

published, it was described as "the largest, most systematic word frequency index" ever compiled (Zeno et al., p. 1).

The word frequency estimates provided by the EWFG, also called standard frequency indices (SFIs), are expressed on a log scale that ranges from a low of 15 (less than .01 occurrences per million words) to a high of 80 (more than 10,000 occurrences per million words.). Words tested on the GRE typically range from a low of 15 (e.g., *epicure*, *conflate*) to a high of 60 (e.g., *level, source*). Since examinees are likely to have the most difficulty understanding the least frequent words, our expectation was that the SFI value of a word, as specified by the EWFG, would be negatively correlated with item delta.

Most word frequency indices, including the EWFG, provide an independent frequency estimate for each possible form of a word. For example, the EWFG estimate for *neutralize* is 39.7, whereas that for its base form, *neutral*, is 52.0. Our analyses suggested that the word frequency estimates obtained for words like *neutralize* tend to overstate the vocabulary knowledge needed for comprehension because they do not take into account the common practice of using information about familiar word forms (e.g., *neutral*) to decode unfamiliar word forms (e.g., *neutralize*). Consequently, the item classification algorithms developed for this study utilized an adjusted word frequency (AWF) estimate, defined as follows:

$$\text{AWF}(word) = \max(\text{SFI}[word], \text{SFI}[\text{base } word])$$

where SFI is the standard frequency index extracted from the EWFG, and base is a function that returns the base form of a word (Deane, 2005). In the example discussed above, the adjustment is evaluated as follows:

$$\begin{aligned}\text{AWF}(neutralize) &= \max(\text{SFI}[neutralize], \text{SFI}[neutral]) \\ &= \max(39.7, 52.0) \\ &= 52.0\end{aligned}$$

As is shown in the following example, the adjustment also accounts for situations in which the base form of a word does *not* facilitate comprehension:

$$\text{AWF}(\textit{incongruous}) = \max(\text{SFI}[\textit{incongruous}], \text{SFI}[\textit{congruous}])$$
$$= \max(36.6, 22.1)$$
$$= 36.6$$

Since each individual VC item includes multiple words, an approach for selecting the individual item words to be included in the word frequency analysis was also developed. The approach involved restricting the analysis to two particular item words: the referenced word in the passage, also called the target word, and the correct option word, also called the key. For the sample VC item shown in Figure 5, these two words are *disingenuously* and *insincerely*, respectively. Although additional item words were considered at earlier stages of the analysis, the results suggested that a strategy of focusing on these particular item words, to the exclusion of all others, would not result in a significant decrease in predictive accuracy.

Figure 7 details the approach implemented to characterize the level of vocabulary knowledge needed to respond correctly to the VC item shown in Figure 5. Three steps are shown:

1. First, the adjusted word frequency approach described above is used to estimate the level of vocabulary knowledge needed to comprehend the targeted passage word (e.g., *disingenuously*).

2. That same procedure is used to estimate the level of vocabulary knowledge needed to comprehend the key (e.g., *insincerely*).

3. These two estimates are compared. The estimate associated with the more difficult word (i.e., the word with the lower AWF value), is taken as the final vocabulary score for the item. Note that this step incorporates the assumption that a correct response to an item is only likely when the examinees' vocabulary knowledge is sufficient to comprehend both the target word and the key.

Step 1: Estimate the level of vocabulary knowledge needed to understand the targeted word in the passage

Voc(*target word*) = AWF(*disingenuously*)

= max(WF[*disingenuously*], WF[base *disingenuously*])

= max(WF[*disingenuously*], WF[*disingenuous*])

= max(15.0, 22.1)

= 22.1

Step 2: Estimate the level of vocabulary knowledge needed to understand the key

Voc(*key word*) = AWF(*insincerely*)

= max(WF[*insincerely*], WF[base *insincerely*])

= max(WF[*insincerely*], WF[*sincere*])

= max(22.1, 46.5)

= 46.5

Step 3: Estimate the level of vocabulary knowledge needed to understand both the target word and the key

Voc(*Item A-2*) = min(Voc[*target*], Voc[*key*])

= min(Voc[*disingenuously*], Voc[*insincerely*])

= min(22.1, 46.5)

= 22.1

Figure 7. **Estimating the level of vocabulary knowledge needed to respond correctly to Sample Item 2 from Passage A.**

Note. Because *disingenuously* does not appear in the *Educator's Word Frequency Guide* (Zeno et al., 1995), the default lower bound of 15.0 is returned.

As noted above, our analysis of the available VC items suggested that, for some items, solution also requires proficiency at understanding a common word used in an uncommon sense. For example, one VC item required understanding the common word *shoot* (AWF = 54.0) used in the sense of *a branch of a family tree.* Similarly, a second VC item required understanding the common word *champion* (AWF = 49.2) used in the sense of *one who acts as a guardian*.

An automated approach for classifying each item as either requiring or not requiring this particular proficiency was specified as follows:

1. If the target word and the key are listed as synonyms in Roget's Online Thesaurus (n.d.), then the item is coded as not requiring the Understand Secondary Word Sense (UndSWS) skill (i.e., UndSWS = No).
2. If the target word and the key are not listed as synonyms in Roget's Online Thesaurus (n.d.), then the item is coded as requiring the UndSWS (i.e., UndSWS = Yes).
3. Note that the sample item in Figure 5 would be classified as UndSWS = No because the target and the key (i.e., *disingenuous* and *insincere)* are listed as synonyms in Roget's Online Thesaurus (n.d.).

Stage 3: Model Development and Validation

The vocabulary features discussed above were evaluated by implementing a regression tree analysis with item delta as the dependent variable and each of the available vocabulary features as independent variables. The resulting tree is shown in Figure 8. In this particular display, each node is plotted at a horizontal location determined from its estimated difficulty value and a vertical location determined from its estimated deviance value. Thus, the horizontal axis tracks changes in predicted difficulty levels as each new variable is added to the model, and the vertical axis tracks changes in the amount of unexplained variation remaining after each new variable is added to the model. The display also shows the number of items assigned to each node (via the node labels) and the individual variables selected to define each split (via the branch labels listed on the edges connecting parent nodes to offspring nodes). The split definitions detail the process whereby the items at each node are divided into smaller and smaller subsets, such that each new subset is defined in terms of a specified combination of features and contains items that are increasingly more homogeneous with respect to observed item difficulty.

The first split is defined in terms of the UndSWS feature. In particular, items coded as UndSWS = No are predicted to be less difficult, and items coded as UndSWS = Yes are predicted to be more difficult. The subset of 16 items classified into the UndSWS = No node are then further split into those with low vocabulary demand (Voc > = 32), which are predicted to be less difficult, and those with high vocabulary demand (Voc < 32), which are predicted to be more difficult. Since the resulting model includes fewer than 10 items in each of its three terminal nodes, additional splits are not feasible and the three-terminal-node solution is accepted as the final model.

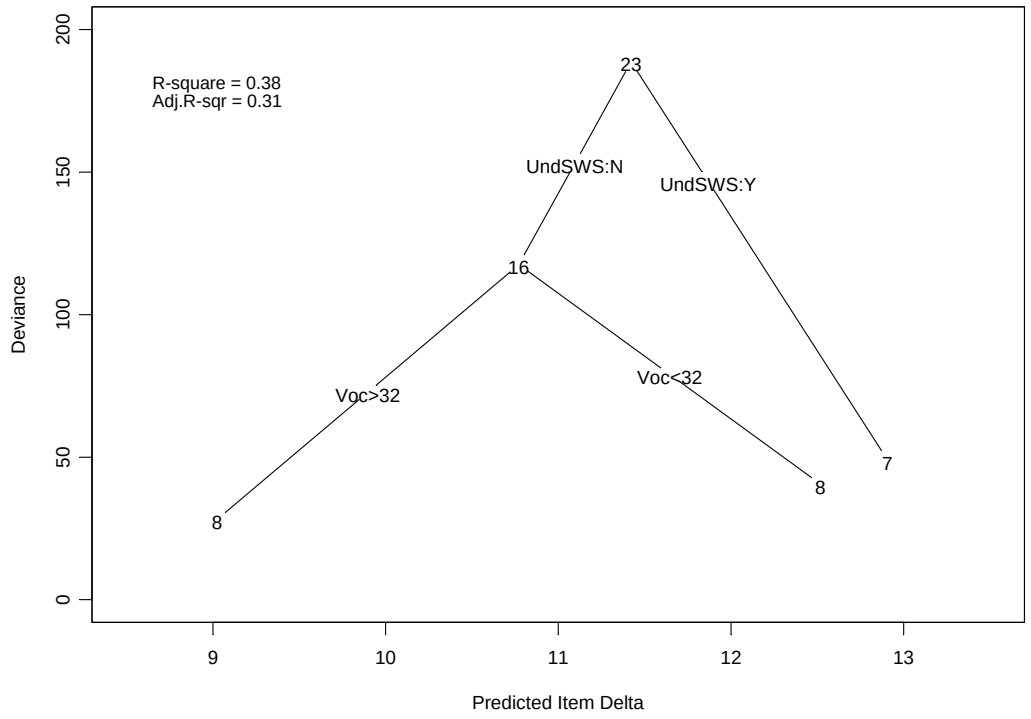


Figure 8. **A tree-based regression model for vocabulary in context (VC) items.**

Note. UndSWS = understand secondary word sense. Items coded as UndSWS:N are predicted to be less difficult, and items coded as UndSWS:Y are predicted to be more difficult.

The regression tree in Figure 8 suggests that the Voc < 32 feature is only significant among items coded as UndSWS = No. In other words, the analysis suggests that word frequency information is only useful for predicting difficulty variation when items do not also require understanding a common word used in an uncommon sense. Table 10 presents a linear regression model designed to test this finding. The analysis confirmed that, as predicted, the Voc < 32 feature is not significant as a main effect but is marginally significant when paired with the UndSWS = No feature ($p < .10$). Thus, both the tree model and the subsequent linear regression model support the finding that items coded as UndSWS = Yes are not appropriately structured for extracting mastery evidence relative to the skill of retrieving critical vocabulary knowledge from long-term memory.

Table 10

Independent Variables, Estimated Regression Coefficients, and Significance Probabilities for Predicting Item Difficulty for Vocabulary in Context Items

Independent variable	Coefficient	se(Coefficient)	*t*	*P*(> *t*)
Intercept	9.0250	0.8649	10.4345	0.0000
UndSWS = Y[a]	4.1917	1.3212	3.1727	0.0050
Voc < 32	–2.1667	2.6424	–0.8200	0.4224
UndSWS = no & voc < 32	5.6604	2.9117	1.9440	0.0669

Note. All independent variables are evaluated automatically.

The preceding analyses yielded detailed information about the cognitive skills needed to respond correctly to VC items located at increasingly advanced levels on the GRE Verbal scale. In Table 11, this information is summarized in terms of a task model for the VC item type. The model lists the individual skills believed to be underlying performance on the VC item type and the particular task features needed to extract mastery evidence relative to each skill. These results are designed to help test developers generate new VC items that provide high-quality evidence about targeted skills and that, as a result, are more likely to scale at targeted difficulty levels.

Table 11

A Task Model for the Vocabulary in Context Item Type

Targeted skill[a]	Required task feature[b]
Understand a high frequency vocabulary word	UndSWS = No & Voc > 32
Understand a low frequency vocabulary word	UndSWS = No & Voc < 32
Understand a common word used in an uncommon sense	UndSWS = Yes

[a]Skills are ordered from easiest to hardest. [b]This column lists the specific task features needed to extract mastery evidence relative to each skill.

Stage 2: Feature Development for IN, PP, and RP Items

This section summarizes analyses designed to explicate the skills underlying performance on IN, PP, and RP items. These three types of items were analyzed as a group because, in each case, examinees' tendencies toward correct response are believed to be critically dependent on proficiency at forming a mental representation of the text that (a) accurately represents the propositions specified in the textbase and (b) incorporates relevant background knowledge.

The analysis considered three different types of task features: (a) features based on the available item format and type classifications, (b) features developed via an automated assessment of the degree of semantic relatedness between the mental representations implied by the subsets of words comprising the passage and each of the available item options, and (c) features assigned via human judgment. These features are described below.

Item type and format features. As noted previously, each item was classified as belonging to one of three different item type categories (IN, PP, or RP) and one of three different response format categories (HS, MC, or SA). For items in the SA category, items were also coded for the number of options classified as correct (i.e., 1, 2, or 3). Two additional item formatting features were also considered: (a) whether the item stem referred to a highlighted sentence in the passage, as in Sample Item 4 in Figure 6, and (b) whether the item stem included a highlighted word or phrase that was also highlighted in the passage. For example, one item that was coded as exhibiting the "Stem Highlights Word/Phrase" feature asked the following question: Which of the following can be inferred about the *university lecture hall* mentioned in the passage? The phrase "university lecture hall" was highlighted both in the item stem and in the passage.

Features developed via an automated assessment of option and passage semantic similarity. Many PR items are designed to gather evidence about examinees' skill at creating a mental representation of the passage that accurately reflects the author's intended meaning. A latent semantic analysis (LSA; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998) was used to distinguish items requiring lower and higher levels of this skill. LSA is an automated text analysis technique that permits quantitative comparisons of the degree of semantic relatedness between any two collections of words. The basic assumptions underlying the approach are that (a) word meanings are accumulated via exposure to large amounts of printed text, (b) important relationships among words may be represented via a finite number of latent dimensions, and (c) the standing of any particular collection of words relative to each of the underlying dimensions

may be determined by considering the patterns of word usage found in a large corpus of relevant texts. Note that these assumptions imply that similarities and differences in word meaning may be quantified by first applying a singular value decomposition to a large matrix of word-by-document frequencies and then representing individual words, and larger segments of text such as sentences or paragraphs, as vectors in the resulting multidimensional space. Landauer and Dumais (1997) have argued that the degree of semantic similarity between any two words, or any two segments of text, such as sentences or paragraphs, then may be approximated by taking the cosine, or dot product, between their respective vector representations.

Many applications of LSA are conducted as follows. First, a large number of documents relevant to the topic of interest are collected, and the semantic relationships within those documents are summarized via a word-by-document frequency matrix. Next, the matrix is subjected to a singular value decomposition analysis, and the resulting *semantic space* is used to develop a vector representation for each text segment of interest. Finally, the degree of *semantic relatedness* among the specified text segments is estimated by calculating the cosine, or dot product, between their respective vectors.

Because the current application is not focused on any particular topic, however, we elected not to develop our own semantic space. Instead, relevant text segments were represented as vectors in the *general reading comprehension space* provided at lsa.colorado.edu. This space was developed from a corpus of more than 37,000 text samples obtained from textbooks, works of literature, and popular works of fiction and nonfiction selected to be representative of the types of reading materials typically considered by students in schools and colleges throughout the United States (Zeno et al., 1995). Previous analyses have suggested that this space provides reasonable estimates of semantic similarity for some purposes (Kintsch, 2002; Wolfe, 2005).

The following procedure was then used to develop a set of automated task features for each of the available items. First, each passage was represented as a vector in the selected space. Next, each item option was also represented as a vector in that space. Finally, a number of cosines were calculated for each item.[4] For the MC items, a total of five cosines were calculated: the cosine between each item's key and the referenced reading passage and the cosine between each of the item's four distractors and the referenced passage. Since $cos(0°) = 1.00$, items that are structured such that the words in the key are nearly identical to the words in the referenced passage should yield key cosines that are near 1.00. Similarly, since $cos(90°) = 0.00$, items that

are structured such that the words in the key are very different from the words in the referenced reading passage should yield key cosines that are near zero. Thus, our expectation was that (a) a low key cosine would be an indication of a *less attractive key* and therefore a *more difficult* item, and (b) a low distractor cosine would be an indication of a *less attractive distractor* and therefore a *less difficult* item.

A single cosine deemed indicative of each item's best distractor was also calculated. This was accomplished by taking the maximum cosine over each of the item's four distractors. This procedure incorporated the notion that, to respond correctly to an MC item, an examinee must be capable of correctly evaluating the truth status of even the most attractive distractor, defined here as the distractor with the strongest link to the passage, as determined from its LSA cosine.

For items written in the SA format, an equivalent procedure was implemented, as follows. First, three cosines were calculated for each SA item, one for each of the three possible choices (i.e., Options A, B, and C). Next, a single cosine deemed indicative of the skill level needed to correctly classify the truth status of each of the item's correct options was obtained by selecting the minimum cosine from among all of the cosines associated with any of the item's correct options. (Recall that SA items may have one, two, or three correct options.) This procedure incorporated the notion that, in order to respond correctly to an SA item, an examinee must be capable of determining the truth status of even the most difficult item option, defined here as the correct option that is *least* closely linked to the passage, as determined from its LSA cosine. Finally, an approach similar to the one described above was used to select a single cosine to represent the single best distractor. In particular, the single best distractor was defined as the incorrect option that yielded the maximum cosine.

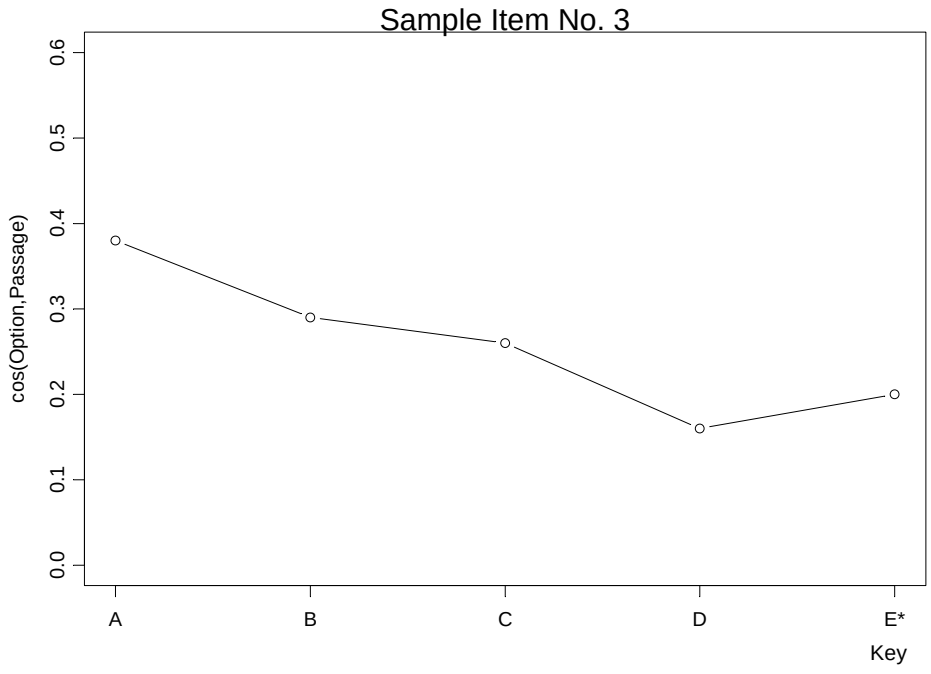
A somewhat similar procedure was implemented for scoring the processing demands of the HS items. First, each sentence in the passage was viewed as a possible correct or incorrect option and was represented as a vector in the selected semantic space. Next, the attractiveness of each sentence was quantified by taking the cosine between its vector representation and the vector representation developed for the item stem. This procedure incorporated the notion that sentences that are more strongly linked to the item stem are likely to be more attractive to examinees, and sentences that are only weakly linked to the item stem are likely to be less attractive to examinees.

Figure 9 displays the LSA cosines obtained for the first two sample items in Figure 6 (i.e., Sample Item No. 3, top plot, and Sample Item No. 4, bottom plot). Note that in each plot the key cosine is relatively low. For example, the key cosine for Sample Item No. 3 is about 0.20, and that for Sample Item No. 4 is slightly lower at about 0.15. These relatively low key cosines suggest that both items are capable of providing mastery evidence relative to the skill of developing a situation model that accurately represents an author's intended meaning when inferential processing is required.

For this item, both the key (Option C) and one of the distractors (Option E) are rated as having a relatively strong link to the passage. When option plausibility is estimated from examinees' observed item response data, however, a slightly different picture emerges. In particular, the observed item response data for this item confirms that, as was predicted by the LSA, a large proportion of examinees selected the key. Among the subset of examinees who failed to select the key, however, the four distractors were about equally attractive. Thus, the high plausibility of Option E predicted by the LSA was not confirmed by the observed item response data.

The plots in Figure 9 also show that, for each item, one or more distractors are rated as being more closely linked to the passage than is the key. In particular, for Sample Item No. 3, Distractors A, B, and C are rated as being more closely linked to the passage than is the key, and for Sample Item No. 4, Distractor D is rated as having this property. This suggests that each item is also capable of providing mastery evidence relative to the skill of evaluating text segments that share a large amount of semantic similarity with a referenced passage but do not provide an accurate representation of the author's intended meaning.

An additional analysis designed to generate more detailed information about the aspects of variation captured by the LSA-based features discussed above was conducted. This additional analysis considered the extent to which option plausibility estimates developed from the LSA-based approach described above agreed with option plausibility estimates developed from an examination of examinees' observed item response data. The analysis indicated that, for certain types of items, LSA-based estimates are not well aligned with estimates developed from examinees' observed item response data. Figure 10 displays the LSA cosines obtained for one such item.[5]



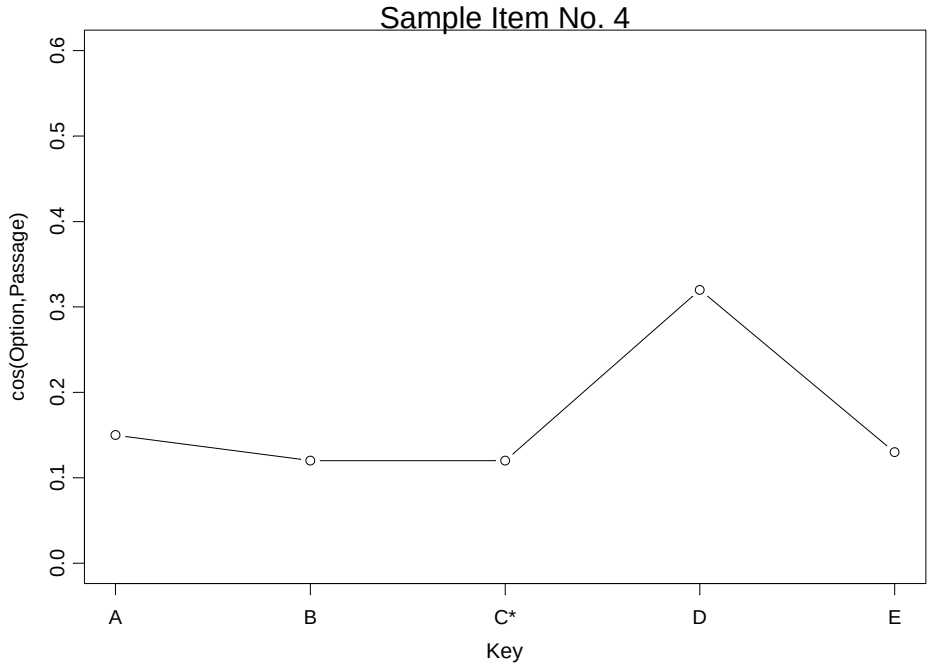


Figure 9. **Latent semantic analysis results for Sample Item No. 3 (top plot) and Sample Item No. 4 (bottom plot).**

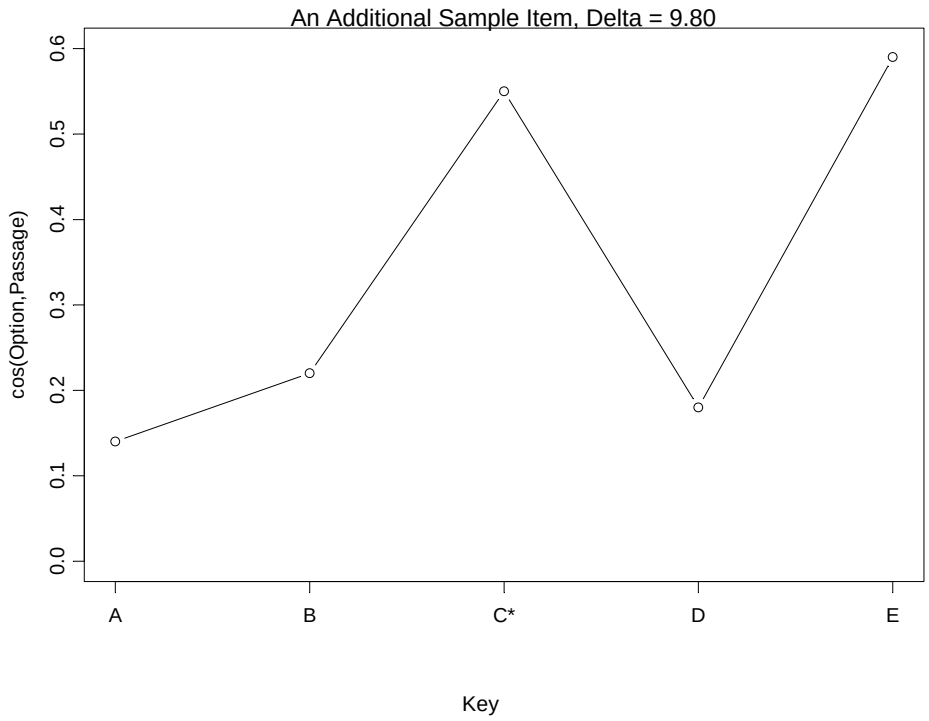


Figure 10. **Latent semantic analysis results for a sample item that contains a distractor that is overtly negated in the passage (Option E).**

Table 12 lists the LSA cosines obtained for a set of five statements selected to shed additional light on these results. Even though Statement 1 expresses an opinion that is directly opposite to the opinions expressed in each of the other four statements, the LSA results indicated a fairly high level of semantic similarity for all pairs of statements. In particular, note that Statement 1, "The argument is valid," and Statement 2, "The argument is not valid," were rated as having the highest possible level of semantic similarity, *cos*(Statement 1, Statement 2) = 1.00. These results illustrate an important limitation of the LSA approach: It is not designed to detect dissimilarities arising from overt or covert negation.

The sample MC item presented in Figure 10 was selected to illustrate this issue. The passage for this item discussed an approach for using sound-absorbing building materials to enhance sound quality. The distractor that was rated as being highly plausible by the LSA, Option E, was specified as follows: "The hall was constructed at a time when sound-absorbing

building materials were not readily available." Thus, although the passage discusses sound-absorbing building materials that *are* readily available, Option E claims that such materials *are not* readily available. This suggests that the high LSA cosine obtained for Option E is an artifact of the fact that the LSA approach is not designed to detect differences associated with overt or covert negation. Also, the fact that Option E was not particularly attractive to examinees suggests that, unlike the LSA analysis, most GRE examinees are adept at discounting item options that are overtly negated in the text

Table 12

LSA Cosines for Pairs of Statements Selected to Illustrate the Effects of Overt and Covert Negation

No.	Statement	1	2	3	4	5
1	The argument is valid.	1.00				
2	The argument is not valid.	1.00	1.00			
3	The argument is invalid.	0.97	0.97	1.00		
4	The argument is suspect.	0.93	0.95	0.97	1.00	
5	The argument is flawed.	0.96	0.96	0.99	0.97	1.00

A feature coded via human judgment. Because the LSA approach described above was found to be insufficient for distinguishing items requiring certain types of oppositional reasoning skills, one hand-coded feature was also included in the analysis. This feature was designed to account for the fact that, as discussed above, many PR passages include extremely complex logical arguments. For example, many passages present arguments in favor of both sides of an issue before revealing the individual side actually subscribed to by the author. Such passages provide an opportunity to gather evidence about an examinee's proficiency at understanding complex oppositional reasoning. One type of item designed to gather this type of evidence is structured such that the argument presented in the key appears to be contradicted by one or more statements from the referenced passage. For example, in one particular PR item, the key was specified as follows: "The author's explanation does not require a revision of scientists' thinking about conditions in the early solar system," even though the referenced passage included the

following seemingly contradictory sentence: "Scientists may have to revise their thinking about conditions at the solar system's formation." Clearly, only those examinees who had closely followed the author's logic would have created a situation model that was precise enough to ensure a correct response to this item. This particular aspect of item variation was captured by defining an additional hand-coded task feature as follows:

Key Opposite Text = Yes if the argument presented in the key appears to be contradicted by a set of propositions in the passage,
= No otherwise.

The validity evidence generated for this feature as well as the additional features discussed above are summarized below.

Stage 3: Model Development and Validation

A correlation analysis was used to provide an initial look at the relationship between item difficulty and each of the specified task features. The resulting correlation coefficients are shown in Table 13. Each coefficient was calculated from a total of 81 items. For each feature, Table 13 also presents a *t* statistic for use in testing the null hypothesis that the corresponding correlation coefficient is equal to zero. These statistics and their corresponding significance probabilities suggested that many of the hypothesized features account for a significant amount of variation in item difficulty. Table 13 also provides an effect-size classification for each feature. These are based on Cohen's (1997) categorization. That is, significant correlations that are at least 0.10 but less than 0.30 are classified as being indicative of a small effect, and significant correlations that are at least 0.30 but less than 0.50 are characterized as being indicative of a medium effect.

Note that six of the features are classified as having yielded a medium-sized effect. Of these, two contributed to decreases in item difficulty and four contributed to increases in item difficulty. The two features that contributed to medium-sized decreases in item difficulty were (a) the HS response format and (b) the stem contains the highlighted word or phrase feature.

The negative coefficient (r = -0.38) and medium-sized effect for the HS response format suggested that items that are formatted such that the response involves highlighting a sentence in the text are likely to be fairly easy. The fact that each text contains just four or five sentences provides a plausible explanation for this result. Note that this result also suggested that the

solution process for items exhibiting the HS response format probably does *not* require development of a detailed situation model. Rather, even those examinees who have succeeded only in creating an incomplete situation model would be likely to respond correctly.

Table 13

Correlations Between Selected Task Features and Item Difficulty for a Set of 81 Inference, Primary Purpose, and Rhetorical Items

Task feature	*N*	*r*	*t*	*p* (> *t*)	Effect size for significant correlations
Response format = highlight sentence	6	-0.38	-3.70	0.0004	Medium
Response format = Multiple choice	55	-0.20	-1.85	0.0675	Small
Response format = select all correct	20	0.45	4.54	0.0000	Medium
Two or more correct options	8	0.30	2.80	0.0063	Medium
Type = rhetorical	15	-0.23	-2.14	0.0351	Small
Type = inference or primary purpose	66	0.23	2.14	0.0351	Small
Stem contains highlighted sentence	8	-0.04	-0.37	0.7098	—
Stem contains highlighted word/phrase	11	-0.34	-3.20	0.0020	Medium
Key opposite text = Yes	9	0.34	3.20	0.0020	Medium
min (*cos*[key, passage))	81	0.09	0.80	0.4273	—
max (*cos*[distractor, passage))	81	0.35	3.30	0.0015	Medium

The HS *response* format should not be confused with the HS *stem* feature, which was *not* significant. Sample Item No. 4 in Figure 6 illustrates this latter feature. The fact that the HS stem feature was not significant suggests that the strategy of using the item stem to reference a highlighted sentence in the text still leaves room for writing item options that are more or less complex, and more or less reflective of the language presented in the passage. Thus, the analysis

suggests that the skills needed to respond correctly to items classified as exhibiting the HS *stem* feature are not significantly different from the skills needed to respond correctly to similarly configured items that do *not* exhibit the HS stem feature.

The second feature that was found to be indicative of a medium-sized decrease in item difficulty is the stem references a highlighted word or phrase in the passage feature. An examination of the 11 items classified as exhibiting this feature indicated that the strategy of referencing a particular word or phrase in the text might have helped examinees to focus their attention on the particular section of the situation model that was most useful for confirming the keys. This suggests that items classified as exhibiting this feature may require a level of inferential processing skill that is lower than that required by similarly configured items that do not exhibit the feature.

Table 13 also lists four features that contributed to medium-sized increases in item difficulty: (a) Format = SA, (b) item has two or more correct options, (c) Key Opposite Text = Yes, and (d) max(*cos*[Distractor, Passage]). These results suggested that item writers might be able to generate more difficult items by (a) using the SA format, (b) including two or more correct options, (c) phrasing the key so that it appears to be contradicted by a sentence in the text, and (d) including at least one distractor that is semantically similar to a sentence from the text.

Additional information about the relationship between these task features and subsequent variation in item difficulty was developed by implementing a tree-based regression analysis with item delta as the dependent variable and the specified task features as the independent variables. The resulting tree suggested that collapsing the available task features is useful, as shown in Table 14. In this particular collapsing strategy, combinations of task features are used to define two skills, each with three different levels (i.e., low, medium, and high): (a) skill at generating required inferences and (b) skill at understanding complex oppositional reasoning. As indicated in Table 14, two or more features are frequently needed to determine an item's optimal skill classification. For example, an item is classified as inferential processing = high if (a) it is formatted as a SA item, or (b) it is formatted as a MC item and *both* of the following are true: cos(key, text) < 0.20 and the stem does *not* contain a highlighted word or phrase.

Table 14

A Task Model for Inference, Primary Purpose, and Rhetorical Items

Skills	Task features
Inferential processing = low	Response format = HS *or* Format = MC & cos(key, text) ≥ 0.50 & max(*cos*[distractor, text))< 0.50
Inferential processing = medium	Response format = MC *and* (stem highlights word/phrase or *cos*[key, text) ≥ 0.20)
Inferential processing = high	Response format = SA *or* (Format = MC & *cos*[key, text) < 0.20 & Stem contains highlighted word/phrase = No)
Oppositional reasoning = low	Response format = HS *or* Response format ≠HS & max(*cos*[distractor, text)) < 0.52 & key opposite text = No
Oppositional reasoning = medium	Response format ≠ HS & max(*cos*[distractor, text) ≥ 0.52 & key opposite text = No
Oppositional reasoning = high	Key opposite text = Yes

The tree-based regression model designed to illustrate the relationship between these skill classifications and subsequent variation in item difficulty is shown in Figure 11. The tree suggests that (a) the easiest items are those that require the lowest level of inferential processing skill and the lowest level of oppositional reasoning skill, and (b) the most difficult items are those that require the highest level of inferential processing skill or the highest level of oppositional reasoning skill. The statistical significance of each split is evaluated in Table 15. For each skill, Table 15 provides a test of the null hypothesis that the specified skill has no effect on item difficulty. Note that all of the main effects are significant, but the specified interaction effect is not. This suggested that the identified skills and their associated task features might help GRE item writers generate new items that are more likely to scale at targeted difficulty levels.

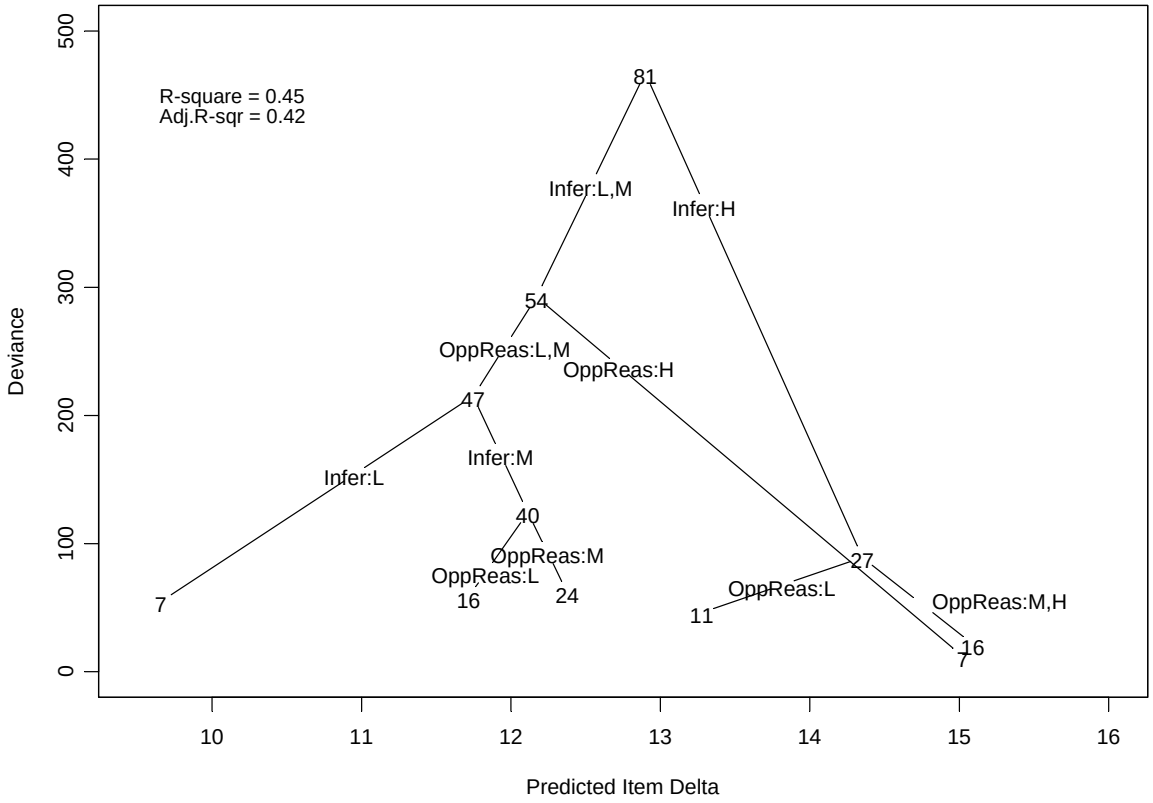


Figure 11. **A tree-based regression model for inference, primary purpose and rhetorical items.**

Table 15

Significance Probabilities for Selected Inference, Primary Purpose, and Rhetorical Features

Feature	Coefficient	se(Coefficient)	*t*	*p* (> *t*)
Intercept	12.55	0.34	36.58	0.0000
Inference = low	-1.96	0.74	-2.64	0.0101
Inference = high	2.14	0.47	4.61	0.0000
Oppositional reasoning = low	-1.04	0.46	-2.28	0.0257
Oppositional reasoning = high	2.76	0.77	3.55	0.0007
Inference = high & oppositional reasoning = high	-0.62	1.54	-1.05	0.2952

Conclusions

The new PR item type is designed to elicit evidence about an examinee's ability to understand and critique complex verbal arguments such as those that are typically presented in scholarly articles targeted at professional researchers. This study demonstrated that a significant amount of the difficulty variance observed for this new item type can be explained by considering item requirements relative to nine particular skills: three vocabulary skills, three inferential processing skills and three oppositional reasoning skills. The study also provided a set of natural language processing tools designed to automatically assess the combinations of skills tapped by individual items. Resulting information was summarized in terms of two different task models: a task model for VC items and a task model for IN, PP, and RP items. The percent of difficulty variance explained by these models ranged from slightly more than 30% for the VC task model to slightly more than 40% for the task model focused on IN, PP, and RP items. These results confirm that the specified task models, together with the associated feature extraction tools, can help GRE test developers develop new PR items that are more likely to scale at targeted difficulty levels. It is important to note, however, that a significant amount of difficulty variation has *not* been explained. This suggests that additional research directed at fine-tuning the models, and the associated feature extraction tools, is also needed.

This study also highlighted a weakness in the automated approach developed to assess items' inferential processing requirements. The approach uses a LSA (Landauer et al., 1998) to evaluate the degree of semantic relatedness between relevant segments of text (e.g., the set of words comprising an item's key and the set of words comprising the referenced reading passage). The analyses confirmed that when text segments involved overt or covert negation, the similarity ratings yielded by the LSA did not agree with those developed from examinees' observed item response vectors. Although a human-coded task feature was developed to account for this in the current study, the results suggested that additional research directed at evaluating alternative approaches for characterizing items' inferential processing requirements is also needed.

Summary, Discussion, and Recommendations

This paper examined alternative approaches for facilitating efficient, evidence-centered item development for the PR item type, an innovative new item type developed for use on the GRE. The results obtained in each of two separate studies are summarized below.

Study 1

The first study described the development and validation of a fully automated natural language processing system designed to help GRE test developers locate the types of stimulus paragraphs required by this new item type. The development approach included two main steps: (a) Human ratings of the acceptability status of a set of candidate source paragraphs were collected, and (b) a natural language processing tool designed to automatically predict those ratings was developed. The validity of the resulting system was evaluated in two separate analyses. First, the paragraph acceptability classifications provided by the human raters (expressed on a 3-point, accept-uncertain-reject scale) were compared to those generated via the automated system. The comparison suggested that the agreement between the automated system and a human rater (on the specified 3-point scale) was nearly indistinguishable from that between two human raters. In particular, the level of exact agreement between the automated system and a human rater ranged from 61% to 62%, depending on the particular subset of ratings considered, whereas that between two human raters was 63%.

The system was also validated by comparing the percentage of acceptable source paragraphs located with and without the automated source-filtering tool turned on. The results of this second validation demonstrated that, when the automated filtering process was *not* used, approximately 10% of the examined paragraphs were found to be acceptable for use in PR passage development. By contrast, when the available paragraphs were first filtered via the specified algorithms, the percentage of acceptable paragraphs located increased to nearly 30%. Since the process of locating acceptable source material is one of the most time-consuming parts of the item development process, this increase should translate directly into efficiency gains.

The algorithms implemented to achieve this increase have been incorporated into the operational SourceFinder system. Consequently, we recommend that the SourceFinder system used to search for new PR source material whenever such material is needed.

Study 2

The development of assessment frameworks that link manipulable task features to the knowledge, skills, and abilities that are the true targets of inference is one of the most important challenges facing test designers today (Mislevy et al., 2002). The second study illustrated one approach for generating and evaluating such links. The approach involved first generating hypotheses about task features that are indicative of increasingly more advanced item processing

requirements, and then using a tree-based regression analysis to evaluate the validity of those hypotheses. The approach was applied to the problem of understanding the knowledge, skills, and abilities needed to respond correctly to PR items with varying stimulus properties. A number of critical task features were identified and linked to specific required skills. Resulting information was summarized in terms of a student model and two different task models. The student model included a total of nine skills: three vocabulary skills, three inferential processing skills, and three oppositional reasoning skills. The associated task models detailed the specific features of tasks needed to extract mastery evidence relative to each skill. The models were validated by considering the percentage of difficulty variance accounted for by the identified task features. This amount ranged from slightly more than 30% for items designed to test vocabulary skills to slightly more than 40% for items designed to test additional verbal reasoning skills, such as generating near and far inferences and understanding complex oppositional reasoning.

These results can be used in a number of different ways. From the test development perspective, the results are likely to be most useful for facilitating targeted, evidence-centered item generation. Our expectation in this regard is that, with sufficient training, GRE item writers should be able to use the models to generate new items that provide more precise evidence about the targeted skills, and that, as a result, are more likely to scale at targeted difficulty levels.

An important limitation of the study, however, is that because the PR item type is a new item type, only limited numbers of items were available for consideration in the analyses. Thus, we recommend that updated versions of the specified models be developed as additional PR items become available.

The detailed information about required skills developed in this study also may be used to describe critical construct elements in ways that may be more illuminating to students, admissions officers, and other GRE stakeholders. Gitomer and Bennett (2002) called this "unmasking the construct" and argued that test designers have an obligation to present such information to test users. The student, evidence, and task models developed in this study provide a straightforward approach for satisfying that obligation

References

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–359). Hillsdale, NJ: Erlbaum.

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

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

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

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

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

Cohen, J. (1997). *Statistical power analyses for the behavioral sciences* (rev. ed.). New York: Academic Press.

Deane, P. (2005, June). *A nonparametric method for extraction of candidate phrasal terms.* Paper presented at the 43rd annual meeting of the Association for Computational Linguistics, Ann Arbor, MI.

Donlon, T. F., & Angoff, W. H. (1971). The Scholastic Aptitude Test. In W. H. Angoff (Ed.), *The College Board admissions testing program*. New York: College Entrance Examination Board.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 300–396.

Embretson, S. E., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343–368.

Enright, M. K., & Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: Implications for item generation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development.* Mahwah, NJ: Erlbaum.

Furnival, G. M., & Wilson, R.W. (1974). Regressions by leaps and bounds. *Technometrics, 16*, 499–511.

Gitomer, D. H., & Bennett, R. E. (2002). *Unmasking constructs through new technology, measurement theory, and cognitive science* (ETS Research Memorandum No. RM-02-01). Princeton, NJ: ETS.

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

Huff, K. (2006, April). *Using item difficulty modeling to inform descriptive score reports.* Paper presented at the annual meeting of the National Council on Educational Measurement, San Francisco.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95*, 163–182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, England: Cambridge University Press.

Kintsch, W. (2002). On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse & W. van Peer (Eds.), *Thematics: Interdisciplinary studies* (pp. 157–170). Amsterdam: Benjamins.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 359–384.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

Larkey, L. S. (1998). Automatic essay grading using text categorization techniques. In *Proceedings of the 21st annual international conference on research and development in information retrieval* (*SIGIR98*; pp. 90–95). New York: ACM Press.

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

Lin, D. (1998) Automatic retrieval and clustering of similar words. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics* (pp. 898–904). Morristown, NJ: Association for Computational Linguistics.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003, July). *A brief introduction to evidence-centered design* (ETS Research Rep. No. RR-03-16). Princeton, NJ: ETS.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development* (pp. 97–128). Mahway, NJ: Erlbaum.

Passonneau, R., Hemat, L., Plante, J., & Sheehan, K. (2002). *Electronic sources as input to GRE reading comprehension item development: SourceFinder prototype evaluation* (ETS Research Rep. No. RR-02-12). Princeton, NJ: ETS.

Powers, D. E. (2000). *Computing reader agreement for the GRE Writing Assessment* (ETS Research Memorandum No. RM-00-8). Princeton, NJ: ETS.

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

Roget's Online Thesaurus. (n.d.). Home page. Available at http://thesaurus.reference.com

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

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

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*, 333–352.

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

Sheehan, K. M., Kostin, I., Futagi, Y., Hemat, R., & Zuckerman, D. (2006). *Inside SourceFinder: Predicting the acceptability status of candidate reading comprehension source documents* (ETS Research Rep. No. RR-06-24). Princeton, NJ: ETS.

Sheehan, K. M., Kostin, I., & Persky, H. (2006, April). *Predicting item difficulty as a function of inferential processing requirements: An examination of the reading skills underlying*

performance on the NAEP Grade 8 Reading Assessment. Paper presented at the annual meeting of the National Council on Educational Measurement, San Francisco.

Sheehan, K. M., & Mislevy, R. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 27*, 1–18.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55*(4), 525–534.

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

Wolfe, M. (2005). Memory for narrative and expository text: Independent influences of semantic associations and text organization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*, 359–364.

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

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide.* Brewster, NY: Touchstone Applied Science Associates.

Notes

[1] We were unable to shift paragraphs from the more populous validation sample to the less populous training sample because (a) the validation sample was drawn after many of the required analyses had been completed, and (b) the validation sample was not randomly selected.

[2] Because rater agreement information was not available for the 47 historical paragraphs, these paragraphs are not included in the summary.

[3] Passage B in Figure 6 is atypical in that it is slightly longer than most PR passages.

[4] As recommended by Kintsch (2002), all cosines were calculated via the document-to-document option.

[5] The text of the item cannot be shown because the item is not disclosed.

GRE-ETS

PO Box 6000

Princeton, NJ 08541-6000

USA

To obtain more information about GRE programs and services, use one of the following:
Phone: 1-866-473-4373
(U.S., U.S. Territories*, and Canada)
1-609-771-7670
(all other locations)
Web site: www.gre.org

* America Samoa, Guam, Puerto Rico, and US Virgin Islands