An Inquiry into the Nature of the Sentence-Completion Task: Implications for Item Generation

Kathleen M. Sheehan
Robert J. Mislevy

July 2001
GRE Board Professional Report No. 95-17bP
ETS Research Report 01-13
An Inquiry into the Nature of the Sentence-Completion Task: Implications for Item Generation

Kathleen M. Sheehan
Educational Testing Service

Robert J. Mislevy
University of Maryland and Educational Testing Service

GRE Board Report No. 95-17bP

July 2001

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

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

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

EDUCATIONAL TESTING SERVICE, ETS, the ETS logos, GRADUATE RECORD EXAMINATIONS, and GRE are registered trademarks of Educational Testing Service.
SAT is a registered trademark of the College Entrance Examination Board.
TSWE is a registered trademark of the College Entrance Examination Board.
TOEFL is a registered trademark of Educational Testing Service.

Educational Testing Service
Princeton, NJ 08541

Copyright © 2001 by Educational Testing Service. All rights reserved.
Abstract

The purpose of this two-part study was to examine the cognitive skills underlying performance on Graduate Record Examinations (GRE®) sentence-completion items. In the first investigation, a database of 112 operationally administered items was used to generate validity information for a subset of hypothesized skills that had been shown to be useful in previous research. Surprisingly, very few of the hypothesized skills performed as expected. In the second investigation, a subset of 45 item variants was used to (a) verify the results of the first investigation; (b) generate new hypotheses about the combinations of item-stimulus features that signal particular required skills; and (c) provide an initial test of those hypotheses. This second investigation verified the findings of the first investigation and provided additional insights into the nature of the sentence completion task. In particular, the analysis yielded detailed information about (a) the mental representations constructed by students during the process of item solving, (b) procedures for determining the subsets of item words that are actually considered by students during the process of item solving, and (c) approaches for predicting the operating characteristics of items when multiple solution strategies are possible. These findings will facilitate subsequent efforts to write new test items that target specific difficulty levels and to create performance summaries that are indicative of students’ strengths and weaknesses.

Keywords: Verbal Reasoning Assessment, Graduate Record Examinations (GRE), cognitive modeling, difficulty prediction
Acknowledgements

This work was supported by the Graduate Record Examinations (GRE®) Research Committee. We are grateful to Adam Banta, Lily Knezevich, and Robin Jacobs for their assistance in obtaining data, coding items, and providing useful information about the combinations of item-stimulus properties that may influence item functioning. We also wish to thank April Ginther of Purdue University for many helpful discussions about the application of psycholinguistic theory to measurement problems in verbal reasoning.
List of Tables

Table 1. Summary of Coded Item-Feature Characteristics ................................................................. 38
Table 2. Output of Leaps-and-Bounds Analysis for Definitional and Reasoning Items.............. 39
Table 3. Within-Family Difficulty Variation for 19 Sentence-Completion Variant Families .. 40
Table 4. Sample Items from Seven Different Sentence-Completion Variant Families ............ 41

List of Figures

Figure 1. Structural templates for three different types of sentence-completion items .......... 43
Figure 2. Scatter plots of calibrated and judged item difficulty for 22 definitional and 90 reasoning items......................................................................................................................... 44
Figure 3. Tree-based regression models developed to explain variation in test developers’ judgements of item difficulty ........................................................................................................... 45
Figure 4. Scatter plots of calibrated item difficulty versus vocabulary difficulty for three different sentence components: key words, stem words, and distractor words ........ 46
Figure 5. Scatter plot of calibrated item difficulty versus sentence length for 112 GRE sentence-completion items ........................................................................................................ 47
Figure 6. An item that requires simultaneous processing of contrasting and causal relations ... 48
Figure 7. Difficulty variation within 19 sentence-completion variant families......................... 49
Figure 8. Item characteristic curves for the parent, one near variant, and one far variant in three different variant families ........................................................................................................... 50
Figure 9. Jump heights observed for four students attempting to jump over a hurdle .............. 51
Introduction

Sentence-completion items appear on the verbal portions of the Graduate Record Examinations (GRE®) General Test and the SAT® I: Reasoning Test. Both examinations use the same basic format: The stem portion of the item presents a syntactically correct sentence in which one or two key words have been replaced by blanks. The option list presents five plausible replacements designed so that all five are syntactically correct but only one is semantically correct.

For the present study, we investigated the nature of the sentence-completion task, paying special attention to (a) the observable item-stimulus properties that signal particular skills required for solving the items and (b) the relationship between these features and resulting variation in the operating characteristics of items. There are important practical reasons for investigating these relationships. Information about the cognitive skills that underlie performance on different types of sentence-completion items – and about the observable item-stimulus properties that signal particular skill requirements – can support needed advances in each of the following areas:

1. Reduction of pretest sample size requirements. As noted in Mislevy, Sheehan, and Wingersky (1993), collateral information about required skills can be used to supplement examinee performance data and allow serviceable estimates of the operating characteristics of items to be developed from smaller pretest samples.

2. Evidence-centered design. A precise understanding of the specific item-feature combinations that signal particular skill requirements can help test development staff design item templates that are optimally configured to yield unambiguous evidence about students’ underlying skill mastery probabilities. (Mislevy, Steinberg, & Almond, 1998).

3. Efficient vat replenishment. Information about the specific item-feature combinations that signal particular skill requirements can help item writers create new items that are targeted at predefined difficulty levels.

4. Automated item generation. When information about the relationship between observable item-stimulus properties and resulting psychometric properties is sufficiently precise, items can be computer-generated as they are administered to examinees. This approach has many advantages, including enhanced security and adaptability (Embretson, 1999).
5. Diagnosis of individual strengths and weaknesses. More accurate information about the skills tapped by individual items can be used to improve diagnosis of skills mastered by individual students. Accurate diagnoses cannot be developed from inaccurate information about these skills.

To the extent that the information provided in this report contributes to a more complete understanding of the relationship between observable item-stimulus features and resulting variation in required skills, it is also likely to prove useful in these applications.

**Review of Relevant Literature**

A number of previous analyses of the skills underlying performance on different types of multiple-choice, sentence-based, verbal reasoning items have been conducted. Five relevant studies are summarized here.

**Item Writers’ Intentions**

DeMauro, Merritt, and Adams (1994) examined the skills underlying performance on GRE and SAT sentence-completion items from a job-analysis perspective. They focused not on individual items, but on the procedures item writers follow to develop those items. After conducting extensive interviews with test development staff involved in the development or review of operational GRE or SAT sentence-completion items, DeMauro and his colleagues compiled a list of the most important written and unwritten rules governing this work. The key results of this research are as follows:

1. Sentence-completion items are designed to sample three different types of semantic relations: definitional, contrasting, and causal. Definitional sentence-completion items are constructed such that the sentence provides a complete definition or example of the key. Contrasting sentence-completion items are constructed such that initial and subsequent portions of the sentence present contrasting or opposite meanings. And causal sentence-completion items are constructed such that initial and subsequent portions of the sentence describe a cause and an effect of that cause.

2. Sentence-completion items are designed to elicit information about an examinee’s ability to determine the particular semantic relation sampled and to select the one option that best preserves that relation. Thus, in every step of the development process, the semantic relations among the words – not the words themselves or the structure of the resulting sentences – are used to determine the final item.
These findings suggest that information about the type of semantic relation considered (definitional, contrasting, or causal) and the type of structure selected to induce those relations may contribute to an enhanced understanding of the knowledge, strategies, and skills involved in solving sentence-completion items.

Insights From an Analysis of Sentence-Correction Items

Bejar, Stabler, and Camp (1987) considered the skills underlying performance on sentence-correction items selected from the now-retired Test of Standard Written English (TSWE®). Two independent measures of syntactic complexity were investigated: sentence length and depth of the parse tree. The analyses confirmed that both measures were significantly related to variation in item difficulty.

While these results are suggestive, their applicability to the current effort is not clear, because TSWE sentence-correction items differ from GRE and SAT sentence-completion items in several meaningful ways. One important difference is that, while sentence-completion stimuli are always syntactically correct sentences, sentence correction stimuli are almost always syntactically incorrect sentences. Secondly, while sentence-completion stimuli are designed to sample a limited range of semantic relations (i.e., definitional, contrasting, and causal), sentence-correction stimuli tend to involve a much wider array of semantic relations. Thus, we should be alert for the strong relationship between item difficulty and syntactic features often observed in verbal tests, but we should also be aware that it may not be present and may not be construct-relevant for GRE sentence-completion items.

Semantic Relationships

Sheehan (1996) investigated alternative graphical approaches for displaying the semantic relations tested by different types of SAT sentence-completion items. Figure 1 provides a sample graphical display. This display highlights the individual words that must be chunked in order to confirm that a given sentence is focusing on a causal, definitional, or contrasting semantic relation. Essential elements of these templates may be described as follows:
1. The causal template distinguishes between the subsets of words comprising the "if clause" and the "then clause."

2. The definitional template distinguishes between the subsets of words comprising the target vocabulary and the sample usage.

3. The contrasting template distinguishes between the subsets of words comprising the initial claim and the contrasting claim.

As noted in Sheehan (1996), comparisons of observable display features may contribute to an enhanced understanding of the structures most often employed to sample definitional, contrasting, or causal relations.

Other Cognitive and Linguistic Features

The skills underlying performance on SAT sentence-completion items were also recently investigated by Buck, VanEssen, Tatsuoka, Kostin, Lutz, and Phelps (1998). A rule-space analysis (Tatsuoka, 1990; Gierl, Leighton, & Hunka, 2000) was used to determine the extent to which sets of hypothesized skills were useful for predicting test takers’ total test scores. The hypotheses tested in these analyses were developed by viewing the sentence-completion item as a mini reading test. The following item features were found to be useful for predicting test takers’ total test scores:

- vocabulary level of words in the item stem, key, and distractors
- word length
- sentence length
- sentence complexity (number of dependent clauses)
- negation in the item stem or key (e.g., presence of "not" or "never")
- number of blanks in the sentence (one or two)
- sentence structure (parallel, opposite, employing cohesive devices)
- sentence content (scientific topics, information density, unfamiliar topics)

The authors concluded that these features are "important determinants of success on the SAT sentence-completion task" (Buck et al., 1998, p.1). It is interesting to note that many of
these same features were considered in several earlier studies of the skills underlying performance on reading-comprehension tasks. For example, Just and Carpenter (1987) studied the effect of sentence negation, and Drum, Calfee, and Cook (1981) studied surface features, such as sentence length, sentence complexity, and vocabulary difficulty. These studies suggest that some or all of the features in this list may help to explain performance variation on GRE sentence-completion tasks.

Insights From Recent Work in Psycholinguistics

Sheehan, Ginther, and Schedl (1999), Sheehan and Ginther (2001, 1999), and Ginther and Sheehan (1999) present a psycholinguistic framework for studying performance on multiple-choice verbal reasoning tasks. The framework builds on the theory of reading comprehension described in Kintsch (1998) and Gernsbacher (1990). Both Kintsch and Gernsbacher argue that reading is a constructive process in which readers attempt to build coherent mental representations of information presented in stimulus materials. Kintsch refers to this process as "construction integration," while Gernsbacher refers to it as "structure building." Kintsch argues that construction integration constitutes a primary goal of comprehension and that efforts directed at satisfying this goal are not related to specific task demands (e.g., instructions to summarize the main idea or identify the topic sentence). Rather, such processing is "an automatic component of the process of comprehension and cannot be separated from it" (p. 174).

Both Kintsch and Gernsbacher attribute individual differences in reading comprehension to individual differences in the mental representations developed by readers during the process of comprehension. Kintsch focuses on two different types of textual representations: the textbase and the situation model. The textbase, as described in Kintsch, is a semantic model of the text (or sentence) and its rhetorical structure. The situation model, on the other hand, is a representation of the text developed by an individual reader. Kintsch notes that these individual mental representations may or may not parallel the actual textbase. This may occur because an individual reader’s mental representation "includes the textbase ... plus varying amounts of knowledge elaborations and knowledge-based interpretations" (p. 50). Kintsch further notes that "[n]either the micro- nor the macrostructure of the situation model is necessarily the same as the micro- nor the macrostructure of the textbase, for the reader may deviate from the author’s
design and restructure the text both locally and globally according to his or her own knowledge and beliefs" (p. 50).

This work suggests that the mental representations that examinees construct and use to solve sentence-completion items need not look like the sentences that are actually presented on the printed page or computer screen. Thus, information about the observable item-stimulus features that trigger differing mental representations may help to distinguish the cognitive skills involved in item solving.

**Method**

The process used to determine the skills that underlie performance on any particular type of item may be described in terms of two functional subtasks: model development and model validation. Model *development* involves the generation of detailed hypotheses about combinations of cognitive skills that could conceivably underlie performance on items with specified stimulus properties. Model *validation* involves the collection of evidence that either supports or refutes those hypotheses.

Alternative validation procedures have also been proposed. In the area of user modeling in cognitive science, for example, detailed production-rule models are used to describe the detailed steps a student undertakes in producing a response, be it correct or incorrect (e.g., Clancey, 1986). To validate these, model-based predictions of correctness and response latency are then compared against results from actual students.

In the current study, validation is accomplished by considering the extent to which hypothesized similarities in required skills are reflected in observed item performance data. The item performance data used for this purpose is the estimated item trace line, or item characteristic curve (ICC), determined from a fitted item response theory (IRT) model.

Why should validation analyses be focused on explaining variation in item trace lines? In many large-scale educational assessments, constraints on item content and pretest psychometric properties are such that items approved for use on final forms tend to exhibit relatively stable
trace lines. The standard explanation for this stability is that the item trace line has not changed because (a) the skills needed to respond correctly to the item have not changed and (b) the distribution of those skills in the examinee population has not changed.

In fact, when significant changes in an item’s estimated trace line are observed, investigations are often initiated to evaluate alternative explanations in terms of a change in the item’s profile of required skills or in the distribution of those skills in the examinee population. This suggests that information about the validity of specific hypothesized skills may be developed by analyzing variation in item trace lines, or more specifically, variation in the parameter estimates that generate those trace lines.\(^1\) In sum, item writers vary features of items in order to focus on different aspects and levels of verbal skill; validity is supported to the extent that these variations correspond to a theoretical rationale of how the operating characteristics of items should be affected, and to empirical evidence about how they actually are affected.

Data Description

A database of 112 operational GRE sentence-completion items was available for consideration. All of these items had been administered on operational paper-and-pencil forms in October 1996. The development, administration, and pretesting of these items did not differ in any way from any other set of operational items administered during this time period. Three-parameter-logistic IRT item parameter estimates, obtained in a series of linked LOGIST calibrations (Wingersky, 1983; Wingersky, Barton, & Lord, 1982), were available for all items. As is the case with all operational GRE LOGIST calibrations, these calibrations provided approximately 1,000 examinee responses for each item. Two independent analyses of the resulting item trace lines were conducted. The design of these analyses are described next.

---

\(^1\) Note that it is highly inappropriate to implement this validation procedure in reverse. That is, it is not appropriate to conclude that two items require similar combinations of skills simply because they exhibit similar trace lines. However, if our proficiency model specifies that two items require the same combination of skills, and those two items do not exhibit similar trace lines, then we have evidence that the assumed model is either inaccurate or incomplete.
Study 1: Analysis of Selected Features and Judged Difficulty Estimates

Study 1 was designed to investigate the predictive capability of 10 specific item features. This set included all features that (a) had been shown to be useful in one or more of the previous research studies summarized earlier and (b) appeared as if they could be efficiently coded. Each feature was evaluated using the full set of 112 items available in the database. Item coding was performed by five members of the Educational Testing Service (ETS®) Assessment Division, each of whom had recent experience either developing or reviewing operational GRE sentence-completion items. Each participant coded one-fifth of the available items. Item statistics were not revealed to staff members either before or during the item coding sessions.

In addition to coding selected item features, these study participants were also asked to provide an item difficulty estimate for each item. These judged item-difficulty estimates were expressed on the ETS delta scale, a nonlinear translation of the percent correct scale.

The coded item-feature data and the judged item-difficulty estimates were analyzed using two different regression approaches: regression by leaps and bounds (Furnival & Wilson, 1974) and tree-based regression (Brieman, Friedman, Olshen, and Stone, 1984). Regression by leaps and bounds provides a comprehensive evaluation of all possible subsets of the independent variables. Since all possible subsets are evaluated, problems that are known to be symptomatic of stepwise procedures are avoided. The tree-based approach provides additional information about the underlying regression relationships (including analyses of interaction effects) and graphical displays designed to enhance substantive interpretations. Both approaches have been shown to be useful in previous studies of item difficulty (Sheehan, 1996, 1997, 2001).

Study 2: Analysis of Item-Variant Families

As part of their operational item development work, test developers sometimes create sets of items that may be viewed as variations on a common theme. These item variants typically share a large number of similar features and a smaller number of distinct features. Item variants may be generated from theoretically derived templates (Enright & Sheehan, in press; Enright,
Morley, & Sheehan, in press) or may be developed in the absence of any preconceived notions about the expected impact of specific item-feature manipulations.

The data set analyzed in Study 1 included 45 items that had been developed as part of an informal variant creation process. These variants were not developed from theoretically derived templates. Rather, individual features of individual items were varied in a nonsystematic fashion to produce 45 unique items that belonged to 19 different variant families of two, three, or four items apiece. Because the number of items involved was not large, the specific feature manipulations that had entered into the creation of each item could be determined after the fact. Study 2 investigated the relationship between these manipulations and subsequent variation in item difficulty. Because the effects of specific item-feature manipulations could be studied given that all other features were held constant, these analyses yielded highly informative results.

Results and Discussion: Study 1

The item features considered in this study included a mix of categorical and continuous variables. Categorical variables included:

- the type of semantic relation sampled (definitional, contrasting, causal, or other)
- the number of blanks to be completed (1 or 2)
- whether the item had a negative – such as "not" or "never" – in either the stem or the key (yes or no)
- the content classification of the item (science, humanities, social science, or arts)
- the structural classification of the sentence (simple, compound, or complex)

ETS test development experts selected these features, in part, because they were confident in their ability to code them reliably. Two continuously-scaled, machine-scored variables were also included: sentence length (number of words in the item stem) and vocabulary level of the item stem, key, and distractors. Following Sheehan (1996) and Buck et al. (1998), the Breland Word Frequency (BWF) indices of the least frequent words in the item stem, key, and distractors were taken as measures of vocabulary difficulty (Breland & Jenkins, 1997).
Table 1 lists the numbers of items classified at each level of each categorical variable as well as the number of items classified within each quartile of each continuous variable. Mean IRT item difficulty estimates for each level and quartile are also provided. As Table 1 shows, all but six of the items were classified as sampling definitional, contrasting, or causal semantic relations. At ETS, items classified as sampling definitional relations are often referred to as definitional items, and items classified as sampling other types of semantic relations are often referred to as reasoning items.

Preliminary tree-based analyses of the available items revealed differing patterns of relationships among definitional and reasoning items. Consequently, all subsequent analyses were implemented separately for definitional and reasoning items. The preliminary analyses also indicated that the data set included two influential items – that is, two items that were located at the extreme ends of the sentence-length distribution. To insure that these items were not allowed to exert undue influence in the estimation of regression coefficients, they were excluded from the leaps-and-bounds analyses reported next. However, these two items were not excluded from any of the other data summaries or plots presented in this report.

Table 2 summarizes the leaps-and-bounds results. For each type of item, definitional or reasoning, four models are listed: the best one-variable model, the best two-variable model, the best three-variable model, and the best four-variable model. Although all possible five-variable and six-variable models were also evaluated, no models containing five or six significant variables were found. The resulting models are highly informative, both because of the coded item features that are included and because of the coded item features that are not included. Important aspects of these models are discussed next.

The Effect of Judged Item Difficulty

As Table 2 shows, test developers’ judgements of item difficulty were useful for explaining variation in item difficulty among definitional items, but they were not useful for explaining this variation among reasoning items. Figure 2 further documents this result using scatter plots of calibrated and judged item difficulty for 22 definitional and 90 reasoning items. These plots include the two items that were found to be overly influential.
As can be seen in Table 2, the available data set provided very few definitional items at the high end of the difficulty scale. However, the handful of definitional items that were, in fact, quite difficult, tended to be judged as such. The scatter plot presented for the reasoning items presents a different story. In particular, reasoning items scaling at the high end of the difficulty scale tended to be judged as having moderate, rather than high, difficulty levels.

Why were these experienced test developers more proficient at estimating the difficulty of definitional items than nondefinitional items? Information relevant to this question was developed by using a tree-based regression technique to investigate the extent to which the test developers’ judgements of item difficulty could be predicted from knowledge of other coded item features. Figure 3 presents the resulting regression trees.

The regression trees depicted in Figure 3 suggest that the judgements test developers provided for definitional and reasoning items differed in terms of the importance assigned to characteristics of keys and distractors. In particular, when judging the difficulty of definitional items, the test developers appeared to be most influenced by characteristics of keys (i.e., vocabulary level and presence or absence of negation). By contrast, their difficulty judgements for reasoning items appeared to be most influenced by distractor characteristics. The relationship between these characteristics and calibrated item difficulty is discussed in the following sections and, later, in Results and Discussion: Study 2.

**The Effect of Vocabulary Difficulty**

As noted earlier, the Breland Word Frequency (BWF) index was used to measure vocabulary difficulty (Breland & Jenkins, 1997). This index was developed from a corpus of 14,360,884 words of running text sampled from reading materials likely to be of use in high-school and first-year college classrooms. According to the authors, the resulting index is believed to be "a pretty good representation of the kind of vocabulary to which high school students and college freshmen are likely to be exposed" (p.1). Several previous analyses of SAT and GRE verbal items have confirmed that, as one might expect, BWF values tend to be negatively correlated with calibrated item difficulty. That is, as measured by the index, verbal items
composed of low frequency words tend to be more difficult, and verbal items composed of high frequency words tend to be less difficult.

The regression models provided in Table 2 illuminate the effect of vocabulary difficulty in the current study. For definitional items, calibrated item difficulty is most influenced by the vocabulary level of the key. (This effect disappears when the test developers’ judgments of item difficulty are included in the model, presumably because they were aware of the role that the vocabulary level of the key can play and factored this information into their judgments.) For reasoning items, the vocabulary level of both the key and the stem are significant.

Figure 4 further documents these findings with scatter plots that show calibrated item difficulty versus BWF value for three different item components: stem, key, and distractors. These plots confirm that the test developers’ expectations regarding the effects of low-frequency vocabulary words among the distractors of reasoning items are not supported by the data.

As Figure 4 indicates, the relationship between calibrated item difficulty and the difficulty of key or stem vocabulary is far from precise. Although key or stem vocabulary difficulty, as measured here, accounts for a significant proportion of the observed variation in calibrated item difficulty, a sizable proportion of total item difficulty variation is not being explained by this item characteristic. These results suggest that either variation in vocabulary demand is not a large factor in determining item difficulty, or current methods for measuring vocabulary demand are not capturing all of the variation in difficulty that is attributable to this factor. As is discussed later, the special nature of the data collected in Study 2, when viewed from the perspective of recent work in psycholinguistics, permitted a much more detailed analysis of the relationship between vocabulary demand and resulting variation in calibrated item difficulty.

The Effect of Sentence Length

Are longer sentence-completion items more difficult to solve than shorter ones? The data collected in the current study suggest that the answer to this question is no. Figure 5 shows a scatter plot of item difficulty by sentence length, which suggests that, considered independently,
sentence length is not at all useful for explaining variation in item difficulty. That is, shorter sentences are just as likely to be difficult as longer sentences. Two extreme cases merit additional attention. As Figure 5 shows, the shortest sentence produced the most difficult item, and the longest sentence produced the least difficult item. The shortest sentence is reproduced here.

The critic accurately described the work as a pastiche, thus recognizing its _____ nature.

(A) allegorical
(B) temperate
* (C) derivative
(D) trendy
(E) realistic

The solution process for this item involves recognizing the definitional relationship between two low-frequency words: "pastiche" and "derivative." Since very little context is provided, all of the information needed to confirm this relationship must be retrieved from long-term memory. This suggests that the extreme shortness of the sentence may be adding to difficulty, not detracting from it, as had previously been suspected.

Additional insights regarding the relationship between sentence length and item solving may be developed by considering performance on the longest item, reproduced here.

By the mid-1980s, most pharmaceutical manufacturers had _____ the exploration of folk uses of plants in their search for new drugs; now, however, the trend has turned back toward _____ the value of plants used in folk medicine as a source of novel therapeutic agents.

* (A) abandoned ... recognizing
(B) begun ... understanding
(C) considered ... investigating
(D) resisted ... ignoring
(E) embraced ... acknowledging
The intended solution process for this item involves recognizing the contrasting relationship between "abandoned" and "turned back toward." However, the fact that this item turned out to be easier than each of the 111 other items studied suggests that many examinees did not, in fact, attempt to reason through the entire sentence. Rather, the extreme easiness of this item suggests that many examinees may have elected to solve it by considering the modified version presented here.

By the mid-1980s, most pharmaceutical manufacturers had _____ the exploration of folk uses of plants.

* (A) abandoned  
  (B) begun  
  (C) considered  
  (D) resisted  
  (E) embraced

Note that any reader who believes that present-day pharmaceutical companies do not advocate "folk uses of plants" would be likely to select the key. Thus, the extreme easiness of this item may also be due, at least in part, to the fact that only one option is consistent with real-world expectations concerning the topic. In addition, the fact that the original sentence was quite long may have even encouraged some examinees to attempt solution by way of the reduced-length, pragmatic approach just described.

This analysis illustrates how knowledge of readers’ underlying goals can help predict the mental representations they construct. In an attempt to use their limited testing time most efficiently, readers may seek to construct a modified representation of the text -- such as the one we provided -- that is "minimally sufficient" to answer the question at hand. It is important to note, however, that readers with different goals -- say, reading to learn or reading for pleasure -- would not be expected to produce the same types of representations.

These results suggest that sentence length, considered independently, is not likely to explain observed variation in item difficulty. But what about the effect of sentence length after all other variables have been controlled for? This is precisely the question that was considered in the leaps-and-bounds analysis. As previously shown in Table 2, even when all possible subsets
of all available item-feature variables are considered, a model that included sentence length as a significant predictor could not be found. Thus, the current data provide no support for the hypothesis that sentence length, measured using numbers of words, is a critical feature in solving sentence-completion items, or for the hypothesis that sentence length is useful for explaining observed variation in the performances provided by different test takers. It is useful to note that these results are consistent with earlier research reported in Klare (1974) and Davidson and Kantor (1982). That is, attempts to artificially manipulate item difficulty by manipulating sentence length were not successful.

The Effect of Other Measures of Sentence Complexity

Three additional measures of sentence complexity were considered: the presence of negatives – such as "not" or "never" – in the item stem or key (yes or no), the number of blanks in the item (1 or 2), and the structural classification of the sentence (simple, compound, or complex.). Negation in the item stem or key was not found to be significant in any of the analyses conducted. That is, even after considering all possible subsets of features, a model that included sentence negation as a significant effect could not be found. However, significant effects were observed for the number of blanks in the item and the structural classification of the sentence. These effects are described next.

Number of Blanks and Resulting Variation in Item Difficulty

The leaps-and-bounds results presented in Table 2 suggest that there is a tendency for two-blank definitional items to be slightly less difficult than one-blank definitional items. The item template shown in Figure 1 offers a plausible explanation for this result. As indicated in the template, definitional items are typically designed to elicit information about an examinee’s comprehension of a particular vocabulary word. This is usually accomplished by presenting a phrase that contains the target word, followed by a phrase that contains a sample usage. When a definitional item has only one blank, that blank is most often placed on the target word. When a definitional item is modified to include two blanks, the first blank tends to remain on the target word, and the second blank is typically placed somewhere in the sample usage phrase. The latter approach is illustrated in the following item.
As if wanting to escape the implications of the sensational charge brought against him, Andrews issued _____ statement claiming that his involvement in the affair had been _____.

(A) a desultory ... provocative
(B) a vindictive ... mischievous
* (C) an exculpatory ... unwitting
(D) an inflammatory ... unpardonable
(E) an exhortatory ... willful

If only one blank had been provided, the solution process would have rested entirely on the examinee’s ability to retrieve the target vocabulary word, "exculpatory," from long-term memory. However, because two blanks are provided, an alternative solution process is available. This alternative process involves deleting the phrase containing the target word, and then proceeding to solve the resulting modified item, as shown here.

As if wanting to escape the implications of the sensational charge brought against him, Andrews claimed that his involvement in the affair had been _____.

(A) provocative
(B) mischievous
* (C) unwitting
(D) unpardonable
(E) willful

The shortened version of the item no longer requires comprehension of the target vocabulary word. Thus, the addition of a second blank has opened the door for solution through a modified sentence that is both shorter and involves a reduced vocabulary demand. Evidence that this alternative solution process is actually taking place is provided by the calibrated item difficulty estimate. This estimate is significantly lower than that of other definitional items with vocabulary demand in the same BWF range as "exculpatory." Thus, the calibrated item difficulty value suggests that the hypothesized solution process is one that is actually being implemented by examinees.
This result is also consistent with the notion that the mental representations constructed by examinees during the process of item solving tend to be minimally sufficient. That is, those stimulus features that are critically needed to select the key tend to be included, while those that are not critically needed to select the key tend to be excluded. Consequently, examinees’ situation models are not necessarily identical to the sentences that actually appear on the printed test form or computer screen.

**Sentence Structure and Resulting Variation in Item Difficulty**

To investigate the influence of structural characteristics, each sentence was classified as belonging to one of the following categories: simple, compound, or complex. As Table 1 indicates, the vast majority of items (78%) were classified as complex. However, information about an item’s structural classification was not at all useful for explaining difficulty variation among definitional items, and it was only marginally useful for explaining difficulty variation among reasoning items. In particular, the analyses confirmed that reasoning items categorized as compound tended to be easier, on average, than simple or complex reasoning items.

In interpreting these results it is important to recall that, in the entire database of 112 sentences, only nine were classified as belonging to the simple category, and only 16 were classified as compound. Consequently, an important finding of the current research is the observation that simple and compound structures are rarely employed in the construction of GRE sentence-completion items. In addition, there is some evidence that completing compound sentences may be slightly easier, on average, than completing complex sentences. However, the small numbers of items available for testing this hypothesis suggests that further investigation is warranted.

**Type of Semantic Relation and Resulting Variation in Item Difficulty**

To investigate possible influences due to the type of semantic relation tested, each sentence was classified as testing one of the following types of semantic relations: definitional, contrasting, causal or other. As Table 1 illustrates, most items were classified as contrasting
(about 40%), followed by causal (about 35%), and definitional (about 20%). Six items were categorized as "other."

Do these different types of semantic relations tend to be associated with items exhibiting specific levels of difficulty? As Table 2 shows, the leaps-and-bounds analysis failed to detect a significant effect due to the type of semantic relation tested. To better understand these results, a small sample of items in each semantic category was investigated further. This second evaluation revealed that, in many cases, two or more different types of semantic relations were tested simultaneously. Consider the item shown in Figure 6. As becomes evident after grouping the phrases into separate semantic chunks, both contrasting and causal elements are present in the item. This suggests that the correct solution requires skill at selecting the set of option words that simultaneously preserves both a contrasting relation and a causal relation.

Are "blended" items – that is, items that simultaneously test two or more different types of semantic relations – more difficult than nonblended items – those that test only one type of semantic relation? Because "blended" was not included as a category in the original data-collection phase of the analysis, the current database does not provide the information that would be needed to precisely answer this question. Additional research in this area is needed.

**Results and Discussion: Study 2**

The goal of Study 2 was to generate additional information about the expected effects of specific item-feature manipulations. The available data included 45 unique items that belonged to 19 different variant families. Table 3 and Figure 7 display the range of difficulty variation found within these variant families. As can be seen, the process underlying the development of these sentence-completion items produced both "near" and "far" variants. *Near* variants are offspring items that preserve parent item parameters. *Far* variants, on the other hand, are offspring items that fail to preserve parent item parameters. Information about those feature manipulations that produced near or far variants can help to distinguish the types of manipulations that do and do not substantially affect cognitive processing requirements. This information, in turn, can contribute to an enhanced understanding of the manipulations that are likely to produce new items that scale at targeted difficulty levels.
Before discussing the types of feature manipulations that do and do not affect item functioning, it is useful to consider the level of variation expected when no features are manipulated – that is, when the same item is recalibrated on a different subset of examinees. Because current item calibration procedures do not yield error-free estimates of item difficulty, items that are recalibrated on different subsets of examinees typically do not yield identical difficulty estimates. The first variant family listed in Table 3 and Figure 7 provides an indication of the magnitude of variation that can be expected when an item is recalibrated on different examinees. This particular family included two items that were identical except for one minor manipulation: The word "his" in the parent was changed to "her" in the child. As can be seen, the resulting difficulty estimates differed by 0.60. This suggests that difficulty estimates that are within 0.60 units of each other may be considered to be statistically indistinguishable – that is, consistent with the same "true" difficulty level.

Table 4, which displays sample items from seven different variant families, illustrates the types of manipulations that did and did not produce statistically distinguishable differences in item difficulty. Specifically, these families were selected to illustrate the types of feature manipulations that yielded near variants – offspring items with item characteristic curves (ICCs) that are nearly identical to parent ICCs – and far variants – offspring items with ICCs that are dramatically different from parent ICCs. In some cases the far variant is easier to solve than the parent; in other cases it is harder. The ICCs obtained for three particular variant families are plotted in Figure 8. The specific feature manipulations employed in the creation of these variants are described next.

**Variant Family 16**

The semantic relation of the parent item for variant family 16 is definitional. As Table 4 shows, the target phrase "affect self deprecation" is defined in the parent as a public stance of "minimizing their own importance." The parent item also presents a reason for affecting this stance: "to avoid accountability." In order to select the words that most accurately convey this particular definitional relationship, an examinee must be capable of understanding the difficult phrase "affect self deprecation."
Variant A presents essentially the same challenge as the parent. The item is still definitional, the target phrase is still "affect self deprecation," and the example or sample usage is still "minimizing their own importance." The main difference between the parent item and variant A is the subject matter of the sentence – that is, its "clothing." Whereas the parent item discusses "political commentators," variant A discusses "presidents of major corporations." Since the parent item and variant A have nearly identical ICCs, these results suggest that such superficial contextual changes do not significantly affect the operating characteristics of GRE sentence-completion items.

Variant B presents a very different challenge. In place of the difficult phrase "affect self deprecation," this variant substitutes a much easier phrase, "deny their own importance." Consequently, the vocabulary level of the variant is much lower than that of the parent. As Figure 8 shows, the ICC of variant B is shifted significantly to the left of the parent ICC. As expected, these results confirm that manipulations of definitional items that fail to preserve the vocabulary level of the target phrase are not likely to preserve parent item parameters.

Variant Family 14

The semantic relation of the parent item for variant family 14 is contrasting. The solution process underlying performance on the parent item involves understanding the contrasting relationship between those who "painted to please their patrons" and Whistler, who "did not eschew innovation." In order to recognize this relationship, an examinee must be capable of understanding the difficult phrase "did not eschew innovation."

Variant A presents essentially the same challenge as the parent. The new solution process involves recognizing a contrasting semantic relation that is described in terms of the difficult phrase "did not eschew innovation." As Table 4 shows, the main difference between the parent item and variant A is the number of blanks. In the parent item, only one word ("innovation") is blanked out, while in variant A, two words are blanked out ("innovation" and "exceptional"). As Figure 8 indicates, however, the parent item and variant A have nearly identical ICCs. This suggests that the number of blanks in a sentence-completion item is not a useful feature to consider when attempting to predict item difficulty.
Variant B presents a very different challenge than that of the parent item. The solution process for this item involves recognizing the contrasting relationship between those who "painted to please their patrons" and Whistler, who did not "conform to the conservative tastes of wealthy benefactors." Even though the difficult phrase "eschew innovation" is still present in variant B, a precise understanding of its meaning is not needed to recognize the intended semantic relation or to select the option that best conveys that relation. Consequently, the vocabulary level of the revised item is much lower than that of its parent; as Figure 8 shows, the ICC of variant B is shifted significantly to the left of the parent ICC. These results suggest that (a) the vocabulary level of a contrasting sentence-completion item is determined from only those words that are needed to convey the intended contrasting relation, and (b) including a difficult vocabulary word in any other portion of the sentence is not likely to affect difficulty. This finding further supports the notion that the situation models considered in item solving tend to be "minimally sufficient." That is, features of the stimulus that are not needed to achieve the goal at hand tend not to be included.

**Variant Family 8**

Like variant family 14, the semantic relation of the parent item for variant family 8 is contrasting. The solution process underlying performance on the parent item involves recognizing the contrasting semantic relationship between a theory that "retained its viability" and one that is "now untenable." In order to recognize this relation, an examinee must be capable of understanding the moderately difficult vocabulary words "viability" and "untenable."

Variant A presents essentially the same challenge as the parent item. The revised solution process involves recognizing the contrasting relationship between a theory that "encountered opposition" and one that enjoyed "widespread agreement that it was sound." In order to recognize this relation, an examinee must be capable of understanding the moderately difficult vocabulary word "opposition" and also the frequent word "sound" used in an infrequent sense. The ICCs for the parent item and variant A are nearly identical. As expected, word substitutions that preserve the semantic relation being tested – and the vocabulary level of the words involved in the expression of that semantic relation – are likely to also preserve parent item parameters.
Variant B presents a very different challenge than that posed by the parent. The solution process for this new item involves recognizing the contrasting relationship between a theory that "retained its currency" and one which is "now moribund." Recognition of this particular semantic relation involves understanding the infrequent vocabulary word "moribund" and a secondary usage of the common vocabulary word "currency." Thus, the vocabulary level of the revised item is much higher than that of its parent, and consequently, its ICC is shifted significantly to the right. These results confirm that word substitutions that preserve the semantic relation being tested, while simultaneously increasing the vocabulary demand of the final item, are likely to produce a significant increase in item difficulty.

The Effect of Selected Item-Feature Manipulations

The detailed discussion just concluded could have been titled, "How to Write Near and Far Variants." Here we present additional support for each of the claims we have presented. We also present and support a number of additional claims.

Clothing changes. Sensibly implemented "clothing" changes do not alter the challenge presented to the examinee. Consequently, this type of contextual manipulation most often results in variant item parameters that are statistically indistinguishable from parent item parameters. The current database includes several instances of variants that differ from their parents only to this extent – such as variant A of variant family 16, discussed earlier. In this case, the parent sentence focuses on "political commentators," while the variant focuses on "presidents of major corporations." In addition, the vocabulary level of the variant is similar to that of the parent: Both items require knowledge of the difficult vocabulary phrase "affect self deprecation." Because the two items are equal with respect to length, vocabulary level, presence or absence of negation, syntactic structure, and semantic relation, any observed differences in the difficulty of the variant would indicate a significant effect due to clothing changes. However, as Table 3 and Figure 8 demonstrate, the item parameters of the variant are statistically indistinguishable from those of the parent. Thus, these data support the hypothesis that clothing changes preserve parent item parameters.
Variant A of variant family 7 provides an additional example of this type of manipulation. In this pair, the parent sentence focuses on "legislation," while the variant focuses on "the four-year plan." In addition, the word "achieved" in the parent is rewritten as "attained" in the variant. Because "achieved" and "attained" are pitched at approximately the same vocabulary level, and the syntactic and semantic structures are unchanged, the item parameters of the variant are statistically indistinguishable from those of the parent (see Table 3). Any observed differences in the item parameters of the parent would have indicated a significant effect due to clothing changes. Thus, these data also support the hypothesis that clothing changes preserve parent item parameters.

Numbers of blanks. The current database includes several instances of variants that were created by manipulating the number of blanks in the parent item stem. Variant A of variant family 14, discussed earlier, provides an illustration of this type of manipulation. This variant differs from its parent only in terms of the number of blanks examinees must complete. As Table 3 and Figure 8 attest, this manipulation did not have a statistically distinguishable impact on item difficulty.

Variant B in variant family 4 provides an additional example of this. Two separate manipulations were implemented to produce this variant. First, the word "intimidated" in the parent was rewritten as "daunted" in the variant; and second, the single blank in the parent was changed to two blanks in the variant. Importantly, the second blank added a somewhat easy vocabulary word – "confidence" – to the item. And as Table 3 shows, the variant was slightly less difficult than the parent as a result of these two changes. Since the substitution of "daunted" for "intimidated" is not likely to have reduced the vocabulary demand of the item, this suggests that the lower difficulty is primarily due to the addition of the second blank. Thus, the data suggest that the addition of a second blank can lessen item difficulty.

However, the reverse of this finding does not appear to be true. That is, the addition of a second blank that involves a more difficult vocabulary word is not likely to contribute to item difficulty, because students are likely to adopt an alternative strategy in which item solution is focused on the easier blank only. Variant A of variant family 11 illustrates this phenomenon. Here again, two separate manipulations were implemented to produce the variant. First, the word
"unsuccessful" in the parent was changed to "failure" in the variant; and second, the variant was written with two blanks while the parent contained only one. In this case the second blank in the variant deployed a relatively difficult vocabulary word, "predominate."

Despite this change in vocabulary demand, the item parameters of the variant were found to be statistically indistinguishable from the item parameters of the parent, confirming that most students opted to solve the item by ignoring the second blank and focusing on the first blank only. Even though the variant included two blanks, as opposed to the parent’s single blank, and even though the variant included a somewhat difficult vocabulary word, a statistically distinguishable increase in item difficulty did not result. These results suggest that attempts to create more difficult sentence-completion items by increasing the number of blanks and attempts to explain low test scores by positing increased difficulty due to an increase in the number of blanks are not likely to be successful.

Manipulations that do not affect vocabulary demand. The current data suggest that the vocabulary level of a sentence-completion item is determined from only those words that are needed to convey the intended semantic relation; including a difficult vocabulary word in any other portion of the sentence will not cause a statistically distinguishable increase in item difficulty. Variant family 14 supports this claim. As Table 4 makes clear, the parent sentence and both variants in this grouping include the difficult vocabulary phrase "eschew innovation." However, as noted earlier, comprehension of the intended contrasting relation posed in both the parent item and variant A requires understanding of the difficult vocabulary phrase "eschew innovation," while comprehension of the intended semantic relation posed in variant B does not require comprehension of this phrase.

Figure 8 clearly demonstrates the validity of this analysis. As can be seen, the ICCs for the parent item and variant A are nearly identical, while the curve for variant B is shifted dramatically to the left. This confirms that even though "eschew innovation" is present in Variant B, an understanding of its meaning is not needed to solve the item; and provides further support for the claim that the vocabulary level of a contrasting item is determined from only those words that are needed to comprehend the intended contrasting relation.
Variant family 9 provides additional support for this claim. In this case, both the parent sentence and variant A are causal. As Table 4 shows, the parent sentence blames "loopholes in government standards" for "one of the most pernicious forms of air pollution," while the variant holds "laxity in the enforcement of safety regulations" responsible for products that are "a menace to consumers." At first glance, one might expect these two sentences to be differentially difficult; although "loopholes" and "laxity" are pitched at approximately the same vocabulary level, "pernicious" and "menace" are pitched at dramatically different vocabulary levels.

However, a precise definition of "pernicious" is not needed to confirm the intended semantic relation. That is, the main message of the parent sentence is that "loopholes" cause "air pollution." Since any kind of air pollution is undesirable, whether or not it is "pernicious" is immaterial. The parent and its variant are structurally equivalent, and this equivalence is reflected in the fact that the two difficulty estimates are statistically indistinguishable. These results suggest that (a) the difficulty level of a causal item is determined from only those words that are critically needed to comprehend the intended causal relation; (b) adding a difficult vocabulary word at any other point of the sentence is not likely to result in an increase in item difficulty; and (c) low test scores cannot be explained by focusing on individual vocabulary words that are not critically needed to confirm the key.

Manipulations that do affect vocabulary demand. Vocabulary demand is affected by certain types of word substitutions and certain types of blank placement modifications. All four far variants in Table 4 and Figure 8 support this claim. In each instance, a difficult vocabulary word was either injected into or excluded from the required solution process. Consequently, in each instance, a statistically significant difference in item difficulty was observed.

Students’ mental representations. As predicted by recent work in psycholinguistics, the current study suggests that the mental representations created by examinees while item solving are not necessarily identical to the test sentences that appear on the printed page or computer screen. All of the variants considered in Study 2 support this claim. In each case, parent/child difficulty estimates were either the same or different as predicted by a psycholinguistic analysis of the types of mental representations that are likely to be created by examinees – bearing in
mind that the readers’ purpose, in each case, is to build a representation that will suffice for answering the question at hand.

Conclusions

The preceding analyses suggest that success at generating sentence-completion items that target specific difficulty levels hinges on advances in each of the following areas:

- designing improved methods for measuring vocabulary demand
- understanding the effects of specific feature manipulations
- detecting and handling items that admit multiple solution strategies
- understanding mental representations constructed by examinees while item solving

Next we discuss our results in terms of each of these areas and summarize the implications for item generation.

Measuring Vocabulary Demand: What Have We Learned?

The preceding analyses demonstrate that vocabulary demand plays an extremely influential role in establishing the difficulty of GRE sentence-completion items. This role is so influential that several other item features appear to be important only to the extent that they establish the subsets of words considered by students while item solving – the words that consequently, need to be passed to subsequent vocabulary assessment algorithms. The current results suggest that future progress at understanding the skills measured by GRE sentence-completion items hinges on our success at developing measures of vocabulary demand that are more reflective of the underlying phenomenon of interest. Some factors which should be considered in this effort are discussed here.

The process of measuring the vocabulary demand of an individual sentence-completion item may be described as follows. First, critical words in the item stimulus are identified. Second, a frequency index is retrieved for each identified word. And third, one or more summaries of those indices are calculated. The vocabulary measures considered in the current
study, and in many previous studies (e.g., Sheehan, 1996; Buck et al., 1998), were based on the null hypothesis that all item words are potentially critical. The current analyses demonstrate, however, that many sentence-completion items contain stem, key, and distractor words that are not considered in students’ solution processes. Consequently, efforts to develop improved measures of vocabulary demand will succeed or fail based on the procedures used to determine the subsets of words that are actually considered in students’ item solving strategies.

The current study demonstrates that two factors are critically important in determining the item words that are actually considered in students’ solutions: (a) the semantic relation being sampled (definitional, contrasting, causal, or some combination of these); and (b) the number and placement of blanks. Both DeMauro et al. (1994) and Sheehan (1996) have noted that different linguistic structures are used to signal definitional, contrasting, and causal relations. Because different structures are used, different procedures may be needed to determine the item words that are actually considered in students’ solutions.

At present, item writers are free to position blanks at any point in the sentence-completion item. However, as demonstrated earlier, small changes in blank placement can result in extremely large changes in vocabulary demand and resulting item difficulty. In addition, different blank-placement strategies are more or less likely to yield items that admit solution through pragmatic approaches that reduce the length of the item. Consequently, guidelines regarding optimal blank placement for definitional, contrasting, causal, and blended items are needed. Such guidelines can help item writers to control the subsets of item words that are actually considered in students’ solutions. As demonstrated earlier, this manipulation can lead to the development of new items that target specific difficulty levels.

**Generating Items With Known Operating Characteristics: How Close Are We?**

Consider Figure 8. In each of these three variant families, we see one variant with operating characteristics that are statistically indistinguishable from those of its parent, and one variant with operating characteristics that are dramatically different from those of its parent. Can we distinguish the types of feature manipulations that produced each variant? The answer, most definitely, is yes. That is, the information developed in the current study is sufficient to create
new variants at predictable difficulty levels. This suggests that, given a sufficient supply of calibrated items, new items with predetermined difficulty levels can be generated tomorrow.

One concern, of course, is whether these new items will look different enough to be useful. The current analyses suggest that some sentence-completion items can be easily reclothed while others cannot. Some experimentation with different types of reclothing strategies is needed before "no data" sentence-completion items can be used operationally.

In addition to detailing the cloning strategies that are likely to yield variant items that inherit parent item parameters, the current analyses provide detailed information about the likely effects of specific item-feature manipulations on newly generated items. These results are particularly notable, since in some cases they are at odds with the prevailing wisdom. Consider the effect of increasing the number of words in a sentence-completion item. Previous research has suggested that, other things being equal, increasing the number of words in a sentence-completion item is likely to produce a more difficult item. The current results are at odds with this prediction. The results obtained in both Study 1 and Study 2 suggest that a strategy of increasing the length of a sentence-completion item either has no systematic effect, or in certain cases, reduces the difficulty of the item.

Previous notions regarding the effect of increasing the number of blanks in a sentence-completion item are also at odds with the current results. In particular, several of the subject matter experts consulted during the course of this study noted that two-blank sentence-completion items were expected to be more difficult than similarly configured one-blank sentence-completion items. The data collected in the current study do not support this hypothesis. In particular, the current results suggest that increasing the number of blanks in a sentence-completion item either has no effect, or in certain cases, produces a less difficult item – never a more difficult one.

Detection and Handling of Multiple-Strategy Items

The preceding analyses demonstrate that current GRE sentence-completion pools contain a fair number of items that are solvable through two or more alternative solution strategies. For
example, as demonstrated earlier, the longest item in the pool could be solved either by reasoning through the entire sentence or by implementing a pragmatic strategy that reduced the length of the item. Similarly, the far variant in variant family 14 could be solved either by understanding the difficult phrase "eschew innovation" or by producing a reduced-length sentence by deleting the phrase containing "eschew." (This second strategy was only feasible for the far variant because of the placement of the blank.) Similar instances of multiple-strategy items were found by Sheehan, Ginther, and Schedl (1999) in their analysis of Test of English as a Foreign Language (TOEFL®) reading comprehension items.

The comprehension models detailed in Gernsbacher (1990) and Kintsch (1998) provide a theoretical explanation for the abundance of multiple-strategy items in current verbal-reasoning item pools. Gernsbacher notes that lower-ability readers are often adept at creating individualized text representations that make sense, even when they do not understand large numbers of seemingly necessary words or phrases. This is accomplished either by deleting unknown words or phrases, or by replacing them with known words or phrases that are not inconsistent with available schema knowledge.

For example, when solving the parent item from variant family 9, one might be tempted to replace "pernicious" with "bad." If the blank is not positioned on "pernicious," this strategy will be highly successful. If the blank is positioned on "pernicious," however, and there is only one blank, then this strategy will not be successful as item writers are not likely to include a common word like "bad" in the option list. As stated earlier, Kintsch (1998) notes in his report that the types of alternative solution strategies identified in this report are "an automatic component of the process of comprehension and cannot be separated from it" (p. 174). Thus, both the current analyses and the psycholinguistic literature support the notion that multiple-strategy items are likely to appear in verbal-reasoning item pools.

Are multiple-strategy items necessarily inappropriate? We would argue that the answer to this question is no. In fact, the ability to implement comprehension strategies that circumvent unknown vocabulary is often considered to be a mark of a good reader. What is inappropriate is expecting the intended solution strategy for a given item to function as the only solution strategy,
and creating models of student proficiency that completely ignore the phenomenon of multiple-strategy items.

Sheehan et al. (1999) present explicit techniques for modifying reading-comprehension items to minimize the probability of solution through alternative, unintended strategies. Similar techniques could be developed for sentence-completion items. However, even if techniques for avoiding multiple-strategy items were developed, test administrators would still need strategies for dealing with the multiple-strategy items present in existing pools. Consequently, approaches for dealing with previously developed multiple-strategy items are also needed.

Given that multiple-strategy items are to be expected and can be detected, how should these items be classified when defining collateral variables for use in predicting operating characteristics of items? While most researchers tend to ignore this issue, others have suggested that it is appropriate to classify such items as if both sets of skills were needed – that is, the set of skills needed to implement the intended solution strategy and the set of skills needed to implement the alternative solution strategy. The current results suggest, however, that this approach can lead to proficiency models that overstate the mastery level needed for correct solution and that overestimate item difficulty.

Figure 9 further explicates the results expected when items admitting multiple-solution strategies are administered to examinees. The figure tracks four hypothetical students attempting to jump over an invisible hurdle. Because these four students are all located at different points on the underlying proficiency scale, each adopts a different jumping strategy, and consequently, each is differentially adept at clearing the hurdle. In particular, the strategies selected by students A and B are inadequate; as a result, their jumps are too low and each fails to clear the hurdle. Students C and D also employ different strategies, but their strategies are successful. Student C jumps just high enough to clear the hurdle. Student D, the most proficient jumper, jumps quite high and clears the hurdle with many extra inches to spare.

Given the results summarized here and illustrated in Figure 9, one might ask, How high is the hurdle? Consider three different methods for answering this question:
• Method 1: Use the jump height recorded for student C only.

• Method 2: Use the jump height recorded for student D only.

• Method 3: Use the average of the jump heights recorded for all students who successfully cleared the hurdle. In this case, this method would yield an estimate equal to the average of the jump heights recorded for Students C and D.

Note that, in the item solving context, these three methods correspond to the following item classification strategies:

• Method 1: Classify multiple-strategy items as if the easiest strategy were the only strategy.

• Method 2: Classify multiple-strategy items as if the most difficult strategy were the only strategy.

• Method 3: Classify multiple-strategy items as solvable through any of the available solution strategies.

Which method is most appropriate? Figure 9 demonstrates that, in the high jump competition, method 1 yields the most accurate estimate of the true height of the hurdle. That is, the best estimate of the true height of the hurdle is not obtained by considering all possible performances that are likely to lead to a successful jump, but rather, by considering only those performances that are likely to just clear the hurdle. Similarly, the analyses summarized in earlier sections of this report confirm that, for the types of items considered in this study, the most precise estimates of item difficulty are not obtained when all possible strategies that are likely to lead to a correct response are considered, but rather, when only that strategy that is just sufficient to provide a correct response is considered. In other words, the most precise estimates of item difficulty are obtained when items are classified using Method 1.

In conclusion, the current research suggests that the students who take our tests tend to create highly predictable mental representations that are, in essence, subsets of original item stimuli. Thus, a reasonable approach to modeling item functioning involves first reducing the item stimulus to that "substimulus" that is actually considered in item solving, and then considering the cognitive processing that is needed to solve that modified item.
References


Tables and Figures
<table>
<thead>
<tr>
<th>Feature</th>
<th>Levela or quartileb</th>
<th>Number of items</th>
<th>Mean difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of Semantic Relation</strong>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrasting</td>
<td>46</td>
<td>-1.20</td>
<td></td>
</tr>
<tr>
<td>Causal</td>
<td>38</td>
<td>-0.81</td>
<td></td>
</tr>
<tr>
<td>Definitional</td>
<td>22</td>
<td>-0.72</td>
<td></td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>Words ≤ 27</td>
<td>30</td>
<td>-0.96</td>
</tr>
<tr>
<td></td>
<td>27 &lt; Words ≤ 29</td>
<td>28</td>
<td>-0.85</td>
</tr>
<tr>
<td></td>
<td>29 &lt; Words ≤ 33</td>
<td>30</td>
<td>-0.94</td>
</tr>
<tr>
<td></td>
<td>33 &lt; Words</td>
<td>24</td>
<td>-1.19</td>
</tr>
<tr>
<td>Number of blanks</td>
<td>Two</td>
<td>83</td>
<td>-1.07</td>
</tr>
<tr>
<td></td>
<td>One</td>
<td>29</td>
<td>-0.72</td>
</tr>
<tr>
<td>Negative in stem</td>
<td>No</td>
<td>69</td>
<td>-1.04</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>43</td>
<td>-0.87</td>
</tr>
<tr>
<td>Negative in key</td>
<td>Yes</td>
<td>18</td>
<td>-1.13</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>94</td>
<td>-0.95</td>
</tr>
<tr>
<td>Sentence structure</td>
<td>Compound</td>
<td>16</td>
<td>-1.23</td>
</tr>
<tr>
<td></td>
<td>Simple</td>
<td>9</td>
<td>-1.00</td>
</tr>
<tr>
<td></td>
<td>Complex</td>
<td>87</td>
<td>-0.93</td>
</tr>
<tr>
<td><strong>Vocabulary</strong>d level</td>
<td>Stem</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BWF ≤ 36.6</td>
<td>28</td>
<td>-0.60</td>
</tr>
<tr>
<td></td>
<td>36.6 &lt; BWF ≤ 42.00</td>
<td>28</td>
<td>-1.01</td>
</tr>
<tr>
<td></td>
<td>42.0 &lt; BWF ≤ 48.15</td>
<td>28</td>
<td>-1.16</td>
</tr>
<tr>
<td></td>
<td>BWF &gt; 48.15</td>
<td>28</td>
<td>-1.14</td>
</tr>
<tr>
<td></td>
<td>Key</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BWF ≤ 36.2</td>
<td>28</td>
<td>-0.59</td>
</tr>
<tr>
<td></td>
<td>36.2 &lt; BWF ≤ 41.0</td>
<td>28</td>
<td>-1.06</td>
</tr>
<tr>
<td></td>
<td>41.0 &lt; BWF ≤ 46.1</td>
<td>28</td>
<td>-1.05</td>
</tr>
<tr>
<td></td>
<td>BWF &gt; 46.1</td>
<td>28</td>
<td>-1.22</td>
</tr>
<tr>
<td></td>
<td>Distractor</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BWF ≤ 29.8</td>
<td>29</td>
<td>-0.83</td>
</tr>
<tr>
<td></td>
<td>29.8 &lt; BWF ≤ 35.5</td>
<td>27</td>
<td>-0.96</td>
</tr>
<tr>
<td></td>
<td>35.5 &lt; BWF ≤ 39.1</td>
<td>28</td>
<td>-0.81</td>
</tr>
<tr>
<td></td>
<td>BWF &gt; 39.1</td>
<td>28</td>
<td>-0.31</td>
</tr>
<tr>
<td><strong>Content</strong></td>
<td>Science</td>
<td>26</td>
<td>-1.32</td>
</tr>
<tr>
<td></td>
<td>Humanities</td>
<td>6</td>
<td>-1.10</td>
</tr>
<tr>
<td></td>
<td>Social Science</td>
<td>41</td>
<td>-1.09</td>
</tr>
<tr>
<td></td>
<td>Arts</td>
<td>39</td>
<td>-0.61</td>
</tr>
</tbody>
</table>

**Note.** BWF = values taken from Breland Word Frequency indices.

a Levels are listed in order of increasing mean difficulty.

b Differences in the numbers of items in each quartile are due to ties.

c Six items were classified as not tapping any of the specified semantic relations.

d Vocabulary level is measured using the Breland Word Frequency value (BWF) of the least frequent word in the item stem, key, or distractors.
Table 2
Output of the Leaps-and-Bounds Analysis for Definitional and Reasoning Items

<table>
<thead>
<tr>
<th></th>
<th>Best one-variable model</th>
<th>Best two-variable model</th>
<th>Best three-variable model</th>
<th>Best four-variable model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definitional items</strong> (n = 21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judged difficulty</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Content = arts</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Blanks = 2</td>
<td></td>
<td></td>
<td>□</td>
<td></td>
</tr>
<tr>
<td>Vocabulary level of stem</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.64</td>
<td>0.60</td>
<td>0.59</td>
<td>0.55</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.25</td>
<td>0.34</td>
<td>0.38</td>
<td>0.45</td>
</tr>
<tr>
<td><strong>Definitional items – excluding judged difficulty</strong> (n = 21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content = arts</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Content = social science</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Vocabulary level of key</td>
<td>▲</td>
<td>▲</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Blanks = 2</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.68</td>
<td>0.66</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.15</td>
<td>0.20</td>
<td>0.30</td>
<td>0.33</td>
</tr>
<tr>
<td><strong>Reasoning items</strong> (n = 89)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary level of key</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Vocabulary level of stem</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Structure = compound</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Content = arts</td>
<td>▲</td>
<td></td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.86</td>
<td>0.84</td>
<td>0.83</td>
<td>0.82</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.11</td>
<td>0.15</td>
<td>0.18</td>
<td>0.20</td>
</tr>
</tbody>
</table>

**Note.** Two overly influential data points were excluded from this analysis. RMSE = root-mean-square error.

▲ = Estimated coefficient is positive and significant at p < .10.
□ = Estimated coefficient is negative and significant at p < .10.
Table 3

Within-Family Difficulty Variation for 19 Sentence-Completion Variant Families

<table>
<thead>
<tr>
<th>Family</th>
<th>Mean difficulty</th>
<th>Parent</th>
<th>Near variant</th>
<th>Far variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2.05</td>
<td>-2.35</td>
<td>-1.75</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-1.98</td>
<td>-1.47</td>
<td>-2.48</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-1.84</td>
<td>-1.33</td>
<td>-2.36</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-1.64</td>
<td>-1.21</td>
<td>-2.07</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-1.58</td>
<td>-1.07</td>
<td>-2.09</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-1.42</td>
<td>-1.86</td>
<td>-0.98</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-1.21</td>
<td>-1.38</td>
<td>-1.04</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.91</td>
<td>-2.04</td>
<td>-1.94</td>
<td>0.79</td>
</tr>
<tr>
<td>9</td>
<td>-0.90</td>
<td>-0.94</td>
<td>-0.86</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-0.87</td>
<td>-0.57</td>
<td>-1.16</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>-0.82</td>
<td>-0.60</td>
<td>-0.60</td>
<td>-1.27</td>
</tr>
<tr>
<td>12</td>
<td>-0.79</td>
<td>-0.97</td>
<td>-0.61</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>-0.78</td>
<td>-0.21</td>
<td>-1.35</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>-0.59</td>
<td>-0.02</td>
<td>0.20</td>
<td>-1.96</td>
</tr>
<tr>
<td>15</td>
<td>-0.53</td>
<td>-0.72</td>
<td>-0.34</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>-0.19</td>
<td>0.34</td>
<td>0.45</td>
<td>-1.36</td>
</tr>
<tr>
<td>17</td>
<td>-0.11</td>
<td>-0.37</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>-0.05</td>
<td>-0.12</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.53</td>
<td>0.24</td>
<td>0.45</td>
<td></td>
</tr>
</tbody>
</table>
### Table 4
Sample Items From Seven Different Sentence-Completion Variant Families

<table>
<thead>
<tr>
<th>Parent item</th>
<th>Variant A: Parent parameters are preserved</th>
<th>Variant B: Parent parameters are NOT preserved</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variant family 16</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political commentators frequently <strong>affect</strong> self deprecation, minimizing their own importance, in order to <strong>avoid</strong> accountability for their opinions, because if their commentary does not matter, no issue of accountability arises.</td>
<td>Presidents of major corporations <strong>affect</strong> self deprecation, minimizing their own importance, in order to <strong>escape</strong> accountability, primarily in order to <strong>disclaim</strong> responsibility for the possible ill-effects of their rhetoric: because if their decisions do not seem to matter, no issue of accountability arises.</td>
<td>Political commentators frequently <strong>deny</strong> their own importance, primarily in order to <strong>disclaim</strong> responsibility for the possible ill-effects of their rhetoric: if their commentary does not matter, the issue of accountability does not arise.</td>
</tr>
<tr>
<td><strong>Variant family 14</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>While most of his contemporaries painted to please their patrons,</td>
<td>In contrast with his contemporaries, most of whom painted to please their patrons,</td>
<td>While most of his contemporaries painted to please their patrons,</td>
</tr>
<tr>
<td>Whistler was exceptional in that he did not <strong>eschew</strong> innovation in order to conform to the conservative tastes of wealthy benefactors.</td>
<td>Whistler was <strong>exceptional</strong> in that he did not <strong>eschew</strong> innovation in order to conform to the conservative tastes of wealthy benefactors.</td>
<td>Whistler was exceptional in that he did not eschew innovation in order to <strong>conform to</strong> the conservative tastes of wealthy benefactors.</td>
</tr>
<tr>
<td><strong>Variant family 8</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The scientist found it puzzling that her rival’s theory retained its <strong>viability</strong> despite widespread suggestions that it was now <strong>untenable</strong>.</td>
<td>The scientist found it <strong>puzzling</strong> that her theory encountered opposition despite widespread <strong>agreement</strong> that it was sound.</td>
<td>The scientist found it puzzling that her rival’s theory retained its <strong>currency</strong> despite widespread suggestions that it was now <strong>moribund</strong>.</td>
</tr>
<tr>
<td><strong>Variant family 4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Despite the fact that his career as a reporter had steeled him to constant contact with the prominent thinkers of his day,</td>
<td>Despite a <strong>confidence</strong> derived from years of experience interviewing the famous performers of her day,</td>
<td>Parkinson was somewhat <strong>daunted</strong> by the prospect of interviewing such a legendary intellectual as Martensen.</td>
</tr>
</tbody>
</table>

*Note.* Spacing between sentence fragments is provided to help in the identification of similarities and differences. These spaces were *not* provided when the items were administered operationally.
## Table 4 (continued)

**Sample Items From Seven Different Sentence-Completion Variant Families**

<table>
<thead>
<tr>
<th>Parent item</th>
<th><strong>Variant A:</strong> Parent parameters are preserved</th>
<th><strong>Variant B:</strong> Parent parameters are NOT preserved</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variant family 7</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The report concluded that the basic aim of the legislation has been <strong>achieved.</strong></td>
<td>The results suggest that the goals of the four-year plan have been <strong>attained.</strong></td>
<td></td>
</tr>
<tr>
<td>or, to be more accurate, that the legislation should not be considered a failure.</td>
<td>or, to be more accurate, that the plan should not be considered a failure.</td>
<td></td>
</tr>
<tr>
<td><strong>Variant family 11</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Since the treaty establishing the European Union called for a common European identity, the Union has been, from this point of view, <strong>unsuccessful:</strong></td>
<td>Since the treaty establishing the European Union called for a common European identity, it is, from this point of view, a <strong>failure:</strong></td>
<td></td>
</tr>
<tr>
<td>national loyalties still predominate among its member countries.</td>
<td>national loyalties are still <strong>predominant</strong> among its member countries.</td>
<td></td>
</tr>
<tr>
<td><strong>Variant family 9</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>To the chagrin of environmentalists, because of loopholes in government standards</td>
<td>Consumer advocates complain that because of laxity in the enforcement of safety regulations.</td>
<td></td>
</tr>
<tr>
<td>one of the most <strong>pernicious</strong> forms of air pollution has become a fact of everyday urban life.</td>
<td>many products on the market are a <strong>menace to</strong> consumers.</td>
<td></td>
</tr>
</tbody>
</table>

**Note.** Spacing between sentence fragments is provided to help in the identification of similarities and differences. These spaces were *not* provided when the items were administered operationally.
Figure 1. Structural templates for three different types of sentence-completion items (displayed items were selected from SAT Form 3QSA01).
Figure 2. Scatter plots of calibrated and judged item difficulty for 22 definitional and 90 reasoning items.
Figure 3. Tree-based regression models developed to explain variation in test developers’ judgements of item difficulty.
Figure 4. Scatter plots of calibrated item difficulty versus vocabulary difficulty for three different sentence components: key words, stem words, and distractor words.
Figure 5. Scatter plot of calibrated item difficulty versus sentence length (number of words in stem) for 112 GRE sentence-completion items.
Surprisingly, given the dearth of rain that fell on the corn crop, the yield of the harvest was *encouraging* and *abundant*; consequently, the corn reserves of the country have not been *depleted* and *extended*.

Figure 6. An item that requires simultaneous processing of contrasting and causal relations.
Figure 7. Difficulty variation within 19 sentence-completion variant families.
Figure 8. Item characteristic curves for the parent, one near variant, and one far variant in three different variant families. In each family, the parent and near-variant curves are nearly coincident while the curve obtained for the far variant is clearly different.
Figure 9. Jump heights observed for four students attempting to jump over a hurdle. Note that the problem of estimating the height of the hurdle is very similar to that of estimating the difficulty of a cognitive test item.