DOCUMENT RESUME

ED 395 032 TM 025 050

AUTHOR	Bennett, Randy Elliot; And Others
TITLE	Toward a Framework for Constructed-Response Items.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-90-7
PUB DATE	Jun 90
NOTE	66p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Classification; *Constructed Response; Models; Multiple Choice Tests; Performance Based Assessment; Responses; *Scoring; Standardized Tests; Test Construction; *Test Items; Validity

ABSTRACT

A framework for categorizing constructed-response items was developed in which items were ordered on a continuum from multiple-choice to presentation/performance according to the degree of constraint placed on the examinee's response. Two investigations were carried out to evaluate the validity of this framework. In the first investigation, 27 test development staff assigned 46 items of various formats to the categories. Overall, agreement with the intended item categorizations was good, with a median of 2 of a possible 27 judges disagreeing with a given item's classification. In the second investigation, responses of 40 examinees each to 4 sets of items were scored by test development staff, with each set scored by 4 individuals. Results showed scoring agreement to be highest for a category requiring the examinee to choose a response from an extended stimulus array and lowest for items requiring that the stimulus be reordered to form a correct sequence. Whether the reported agreement levels represent sufficient accuracy to permit the widespread use of such items in standardized tests depends on whether some degree of scoring error, however small, can be accepted. Appendix A gives sample items organized by category, and Appendix B is a scoring guide organized by item category. (Contains 1 figure, 7 tables, and 13 references.) (Author/SLD)

RR-90-7

RESEARCH REPORT

TOWARD A FRAMEWORK FOR CONSTRUCTED-RESPONSE ITEMS

Randy Elliot Bennett
William C. Ward
Donald A. Rock
Colleen LaHart

Educational Testing Service
Princeton, New Jersey
June 1990

Toward a Framework for Constructed-Response Items

Randy Elliot Bennett

William C. Ward

Donald A. Rock

and

Colleen LaHart

Educational Testing Service

Acknowledgements

Numerous people contributed to this project. Peter Cooper coordinated a team of ETS test developers responsible for item and scoring guide development; Philip Oltman, Gordon Hale, and Ann Gallagher pilot tested scoring keys; Hazel Klein and Terri Stirling managed data collection; and several dozen test developers categorized and scored items. The students and staff of Bunker Hill Community College (MA), Central Piedmont Community College (NC), the College of the Desert (CA), Santa Fe Community College (NM), and Lewis and Clark Community College (IL) responded to test items to provide data for investigating scoring reliability. Finally, Michael Zieky and Mari Pearlman offered thoughtful reviews of an earlier draft of this report that, along with the experience of this project, have helped us better appreciate the considerable complexity of constructed-response testing.

Abstract

A framework for categorizing constructed-response items was developed in which items were ordered on a continuum from multiple-choice to presentation/performance according to the degree of constraint placed on the examinee's response. Two investigations were carried out to evaluate the validity of this framework. In the first investigation, 27 test development staff assigned 46 items of various formats to the categories. Overall, agreement with the intended item categorizations was good, with a median of two of a possible 27 judges disagreeing with a given item's classification. In the second investigation, responses of 40 examinees each to four sets of items were scored by test development staff, with each set scored by four individuals. Results showed scoring agreement to be highest for a category requiring the examinee to choose a response from an extended stimulus array and lowest for items requiring that the stimulus be reordered to form a correct sequence. Whether the reported agreement levels represent sufficient accuracy to permit the widespread use of such items in standardized tests depends on whether some degree of scoring error, however small, can be accepted.

Toward a Framework for Constructed-Response Items

The multiple-choice item has been and remains the mainstay of large-scale testing programs in the United States. There are several reasons that support this choice. First, compared with item types requiring judgmental keying, scoring is objective and reliable. Moreover, test scoring can be automated and thus can be inexpensive and swift. Third, relative to some other formats, items can be answered very rapidly. This means that, within a limited period, it is possible to obtain the broad content sampling necessary to assure that a test provides a reliable and generalizable representation of a domain. Finally, a sophisticated statistical technology has been built to support the analysis of these items (e.g., Lord, 1980).

Whereas multiple-choice items have important advantages, they also have significant limitations (N. Frederiksen, 1984). For one, multiple-choice items are more easily used to test specific, isolated pieces of knowledge than to measure higher-order skills, such as problem solving, in the real-world contexts in which they are normally used. Second, these items can be answered correctly, with relatively high probability, by guessing. Guessing introduces error into the measurement of performance, particularly for low ability examinees and for difficult tests. Third, unless items are very carefully constructed, this item type is susceptible to coaching based on strategies that deal with superficial characteristics of the item rather than the examinee's knowledge of the content the item is intended to assess. Finally, as usually constructed and scored,

the multiple-choice item does not provide for the assessment of partial knowledge or for the identification of diagnostic information concerning the source of an examinee's errors. While there are techniques by which this limitation can be overcome (e.g., Coombs scoring to assess partial knowledge), other item formats may be more suitable for the attainment of these objectives.

The limitations of the multiple-choice format become particularly evident when viewed in the context of recent pressures for educational reform (Fiske, 1990; J. R. Frederiksen & Collins, 1989). In this context, tests are expected to (1) emphasize higher order processes so that problem-solving skills will be more rapidly incorporated in curricula, (2) facilitate instruction by identifying specific skills individual learners have yet to master, and (3) measure the outcomes of curriculum reform efforts intended to enhance higher-order skills.

Various item formats retain the amenability to machine scoring of the multiple-choice item while ameliorating some of its less desirable features. Carlson (1985) describes a number of these formats. Particularly attractive are variations of the keylist or master-list item. In one version, a set of item stems is presented with a common list of possible responses; correct responses to one stem serve as distractors for the others. This format eliminates the need to create plausible distractors for each stem. The use of a relatively long list of possible responses can reduce the probability of correct guesses and of "gamesmanship" strategies for choosing correct answers. The

format can also allow for multiple correct responses to a question in order, for example, to accommodate regional differences in terminology.

Whereas item types like the keylist can increase the flexibility with which assessment is performed, they are not sufficiently open for some assessment purposes. The limitation to items in which the examinee is to recognize a correct option is artificial; some real-world situations have this character, but others require that an individual generate solutions to a problem without being presented with the alternatives (Ward, N. Frederiksen, & Carlson, 1980). Still others require that the individual identify the problem, rather than address a problem posed by someone else. Tests that mirror these characteristics of skilled or intelligent performance are needed to provide valid representation of the range of skills for which assessment is desired (Nickerson, 1989; J. R. Frederiksen & Collins, 1989). Such measures are particularly relevant when the interest is in assessing higher-order skills--the ability to acquire, organize, and apply knowledge and strategies--rather than the simple possession of information or algorithms.

Over the past decade, educational researchers in reading, writing, and mathematics have increasingly emphasized the need for instruction in the thinking and problem solving skills required for competence, as opposed to a concentration on mechanics and errors. If for no other reasons than to secure credibility and face validity, assessors may need to provide

instruments that involve significant productivity on the part of examinees.

With the prospect of changing needs and uses for test information, and with the increasing availability of technologies that can facilitate the scoring of more complex responses (Bennett, in press), it is appropriate to rethink our dependence on the multiple-choice item and consider the advantages and limitations of potential alternatives. That process should be aided by a framework for organizing item types. Such a scheme should help identify relevant item characteristics, suggest research questions, aid in organizing research results, and perhaps stimulate new development directions.

This paper presents the beginnings of such a framework by describing an initial set of item categories intended to capture the range of constructed-response item types. Also reported are empirical analyses of the consistency of judges' classifications using these categories and of the relationship between category membership and scoring reliability.

A Preliminary Framework

Figure 1 depicts a categorization of item types according to the task presented, where the main variant is the extent of openness allowed in the response. The categories, which were constructed from a review of individually and group-administered achievement and ability test items, are intended to represent discernible points along this "openness" continuum. Seven categories are listed from more to less constrained: multiple-choice, selection/identification, reordering/rearrangement,

substitution/correction, completion, construction, and presentation/performance.

Insert Figure 1 about here

To be of minimum utility, such a categorization must have at least two characteristics. First, classifications of items into the intended categories must be consistently made by any reasonable judge. Second, categorizations must be associated with item attributes deemed important to the measurement process. One such property is scoring objectivity. At the extremes, the objective nature of multiple-choice items is well established whereas experience suggests that presentation/performance tasks are more difficult to grade reliably.

This report presents data on both characteristics of the categorization. Specifically, the concern was with (1) whether independent judges agreed among themselves in placing items into the intended categories and (2) whether judges could score items from different categories with equal accuracy.

Method

Judges

Judges were two groups of ETS test development staff experienced in the construction of verbal tests. For the first part of the study (assessing the consistency of classifications), fifty-three test developers were asked to participate and 27 returned responses. Of those participating, the mean number of years of test development experience was 10.0, with a standard

deviation of 6.4. Just over half of these individuals (52%) reported as their highest degree a master's with most of the remainder (40%) holding the doctorate. Most individuals (68%) indicated that the humanities constituted their major field of study, with all but one other majoring in the social sciences.

For the second part (assessing scoring objectivity), 16 test developers were sought. The 16 who agreed to participate had a mean of 9.8 years of test development experience (standard deviation = 7.1). Nine reported as their highest degree a master's with all but one of the rest holding the doctorate. Again, most (12 of 16) indicated that the humanities constituted their major field of study, the others having had their formal education in the social sciences.

Procedure

A third group of nine test developers experienced in the generation of verbal tests was asked to construct multiple items conforming to the specifications for categories 0-5 above, with more than one item subtype represented in each category. Each developer was asked to write items that measured sentence-level verbal skills (e.g., grammaticality and style, semantic processing) at a level appropriate to college freshmen. Because these skills could not easily be represented using presentation/performance items, this category was dropped from the empirical investigation. Developers were directed to write items measuring, to the extent possible, similar content within and across item categories, with the ideal result being items

distinguishable from one another primarily on the basis of response format.

In all, 46 items were developed. Because the category definitions given test developers left room for interpretation, the items did not in all cases fit the categorizations intended by the authors. For this reason, several items were reassigned. The resulting distributions ranged from 6 to 9 items per category (see Appendix A).

Item classification. To explore the consistency and correctness of item classifications, test developers not involved in the construction of the items (the first group of judges described above) were given the category specifications and asked to classify each item without knowing to which category the item belonged. Consistency was estimated using the models and methods of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In generalizability theory, variances associated with the different components contributing to the total variation in a set of test scores, or ratings, are estimated. These variance components are assigned to true or error variance depending upon the purpose of the measurement procedure.

A three-way analysis of variance was used to estimate the variance components of the following mixed model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + I_{k(i)} + \alpha\beta_{ij} + \beta I_{jk(i)}$$

where Y_{ijk} is the ordered classification assigned by the jth judge to the kth item in the ith category, α is the category effect, a fixed facet representing the complete population of categories,

and β, judge effect, and I, item effect, are random facets presumed to be sampled from infinite populations of judges and items, respectively. The data analyzed were the classifications assigned by each of the 27 judges for six items sampled from each of the six categories.

For this analysis, category was considered to represent true variance. Allocated to observed variance was, in addition to category, variance due to judges, to items within categories, to the category-by-judge interaction, and to the item-by-judge within category interaction. A generalizability coefficient for a single judge was computed by dividing the true variance estimate by the observed variance as per Thorndike (1982, p. 166-167).

To explore the correctness of item classifications, several analyses were conducted using all 46 items. First, the product-moment correlation between each judge's classifications and the item's intended category classification was computed and these correlations averaged using the Fisher r-to-z transformation.

Because this analysis is sensitive only to the extent to which judges order the items similarly to the intended ordering, and not whether the category placements themselves are correct, the difference between each judge's categorization for an item and the intended categorization was computed and averaged across items. This mean signed difference indicates the extent to which the judge misclassified items, even if the ordering was similar to the intended one, and in what direction the judge's classifications diverged.

Finally, the number of judges diverging from the intended categorization for each item was computed and averaged across items within a category to identify which categories were the most difficult to classify. The frequency of disagreements for each item was also examined to identify problematic subtypes.

Scoring reliability. To evaluate the relationship between category membership and scoring reliability, responses to 42 of the 46 items were collected from student volunteers attending Bunker Hill Community College (MA), Central Piedmont Community College (NC), the College of the Desert (CA), Santa Fe Community College (NM), and Lewis and Clark Community College (IL). (The four essays were eliminated to keep the test from consuming an inordinate amount of examinee time.) The 42 items were divided into four approximately parallel forms (A-D), of 10-11 items and each form was administered to a random quarter of the students at each college. From the 212 completed student tests, 160 were randomly chosen and divided into four sets of 40. Each set of 40 tests was given to a different group of four raters for scoring. The raters independently scored the tests according to a scoring guide (see Appendix B) that was reviewed, pilot tested, and revised before use.

For each item, the scores of the four raters for the 40 examinees were subjected to variance components analysis using the following model:

$$Y_{ij} = \mu + \pi_i + \beta_j + \pi\beta_{ij}$$

where Y_{ij} is the item score assigned by the jth judge to the ith examinee, π is the person effect, and β, the judge effect, with

both effects presumed to be random. Considered as error were the judge effect and the person-by-judge interaction.

A generalizability coefficient for a single rating was computed for each item along with the median coefficient for each category. Emphasis in interpreting these coefficients was on their relative rather than absolute magnitude. This emphasis was chosen because, although care was taken in developing the scoring guide and communicating the task to judges, this experimental grading lacked the protections and motivations characteristic of operational free-response scorings, such as those conducted by the College Board's Advanced Placement Program (Jensen, 1987). To illustrate, raters in these sessions commonly spend almost as much time training as the total time available for our experimental scoring. Second, the guides developed are often critiqued and refined as part of this extensive training process. Third, the graders' performance is monitored in real time to detect and correct misapplications of the scoring guide. Finally, raters are motivated by the fact that their judgments might have an important impact on the examinee. As a consequence, this study likely underestimates the agreement levels that might be obtained under these more stringent, operational conditions.

Because the categorization scheme orders items by the openness of the response and because greater openness is generally associated with lower scoring reliability, item category membership should be related to scoring reliability. To test this hypothesis, items were collapsed across forms and the

Pearson product-moment correlation between each item's generalizability coefficient and its category membership was computed. The significance of this correlation was tested using a one-tailed *t*-test with alpha set at .05.

Finally, to assess the extent to which multiple ratings might allow the generalizability coefficients of the constructed-response categories to approximate the agreement levels of multiple-choice, the single-rater coefficients described above were stepped up using a method suggested by Winer (1971, p. 287). Coefficients were generated for the mean ratings of two, three, and four judges.

Results

Item Classification

Results of the variance components analysis are shown in Table 1. As can be seen, the largest variance estimate is for category, the object of classification, which is true variance. Among the error components, the overwhelming portion of variance is associated with the item-by-judge within category interaction. This interaction indicates that for any given category, the extent of agreement among judges differs as a function of the particular item being classified. The single-judge generalizability coefficient for these data is .95, suggesting that classifications can be reliably generalized across judges and items.

Insert Table 1 about here

Table 2 presents the distribution of product-moment correlations between the intended item categorizations and the categorizations made by each judge for all 46 items. As the table shows, most judges were able to reproduce the rank ordering of the intended categorizations reasonably well: only four of the 27 correlations fell below .85 and the mean and median correlation were .94 and .93, respectively.

Insert Table 2 about here

To determine whether judges' classifications simply tended to duplicate the rank order of the intended classifications as opposed to the actual placements, the intended category designation for an item was subtracted from the judge's designation and these differences averaged across items. The distribution of these mean signed differences is presented as Table 3. As the table indicates, judges' categorizations deviated little from the intended ones: the mean and median of this distribution were .11 and .07, respectively, an average deviation of a fraction of a category per item per judge.

Insert Table 3 about here

Examination of the average number of judges diverging from the intended item categorization--where the range of disagreements for an item is 0 to 27--gives an indication of which categories were the most problematic. Median disagreements

were highest for the selection/identification category (Md = 6) and lowest for construction and reordering/rearrangement (Md = 0 and .5 , respectively). The remaining three categories--completion, multiple-choice, and substitution/correction--fell in between with median disagreements of 2, 2.5, and 3, respectively.

Shown in Table 4 are the number of judges diverging from each item's intended categorization. This table identifies what items stand out within categories as difficult to validly classify. The category with the largest median number of disagreements, selection/identification, is represented by two item types, cloze elide and keylist. Neither type seems more prone to disagreement than the other. From a review of the judges' classifications, these items most frequently appeared to be confused with completion items (the case for two of the three keylists) or with substitution/correction (the case for all five cloze elide questions). Two of the three keylist items do, in fact, take a sentence completion format (but one that is followed by a list of alternative words to be used to complete the sentence), accounting for some judges' confusion. (This same confusion was evident for item #13, a multiple-choice question presented in a sentence-completion format.) The cloze elide items present passages which contain irrelevant or incorrect words that the examinee is asked to strike out (i.e., select). The confusion here seems to be that the examinee is <u>correcting</u> the passage, which makes the item appear superficially appropriate for the substitution/correction format.

Insert Table 4 about here

Several other items warrant discussion because of the high levels of disagreement found for them. The item most difficult to classify belongs to the substitution/correction category. This "construction shift" task (item #41) requires the examinee to rewrite a sentence given a new beginning so that the surface structure is changed but the original meaning is preserved. This item was most often misclassified as reordering/rearrangement, an understandable choice given that the task appears to be to simply rearrange the stimulus sentence. However, the task is slightly more complex as the change in structure typically requires some modification of the original beyond rearrangement. This modification may include changes in verb forms or the addition of connectives. As a result, the task is arguably one of substitution, though some amount of rearrangement does play a part.

The next two most disputed items were completion questions. The word insertion task (item #3) asked the examinee to insert words into an incomplete sentence to make the sentence logically and grammatically correct. Item #35 is conceptually similar, requiring incorporation of appropriate punctuation. In both cases, the items were most commonly mistaken for members of the substitution/correction category, probably because the stimuli were to be corrected and no blanks--which are commonly associated with the completion format--were included.

Scoring Reliability

The results of the analysis of scoring agreement are presented in Table 5. Shown are generalizability coefficients for each item (ordered by category within form), where each form of 10-11 items was scored by a different group of four raters. These coefficients reflect the level of reliability that would be obtained from a single reading with level differences among raters included as measurement error. Primary interest is on the relative differences among items as opposed to the absolute reliability levels. The median coefficients for Forms A-D were .87, .85, .67, and .73, respectively. Taking the bottom third of the coefficients in each form, some consistencies are evident. First, reordering/rearrangement and substitution/correction items are among the ones with the lowest coefficients. In three of the four forms, completion items also appear in this group. Several subtypes appear several times each (e.g., word rearrangement, sentence combining, and sentence completion), though the presence of single instances of several subtypes in the item set suggests that this consistency be cautiously judged.

Insert Table 5 about here

To assess the relationship between item category and scoring reliability, the product-moment correlation between category membership and generalizability was computed collapsing the items across forms. The relation was as predicted, though moderate at -.36 (t = -2.38, df = 40, p < .05).

Another view of the relationship between category and scoring reliability is presented in Table 6, which shows the median and range of generalizability coefficients by category, again with the coefficients collapsed across forms. The highest medians (.93 and .87) and narrowest ranges are associated with the multiple-choice and selection/identification categories. Reordering/rearrangement and completion evidence the lowest medians (.56 and .67), and completion and substitution/correction the widest ranges.

Insert Table 6 about here

Some insight into the causes of disagreement for particular items can be gained from an informal look at the scores assigned by the raters. In several cases, the data suggest that low agreement levels could be attributed to a single rater (though not always the same single rater). In some instances, a rater appeared to misunderstand the allowable range of scores, perhaps from having to repeatedly switch scales from item to item. On form B item #16 (sentence ordering), one of the raters awarded scores to a quarter of the examinees that were beyond the range of the scale. On form C #46 (sentence combining), one rater graded all papers on a 0-1 scale instead of the indicated 0-3 scale. In this case, if read too quickly the scoring guide might be taken to imply that two alternative scoring schemes existed (because of the placement of a capitalized "OR"). When the scores for these single raters are removed, the agreement

coefficients change from .53 to .76 for item #16 and for item #46 from .30 to .61 (where the new coefficients are based on three rather than four raters).

Individual raters also appeared occasionally to diverge from the group because they applied the guidelines for what constituted a correct answer more or less strictly. On form D #40 (a word rearrangement task), one of the raters consistently gave credit to a greater range of responses than the key allowed, presumably because the rater believed the added responses to be correct. Removing this rater's scores resulted in an increase in the generalizability coefficient from .52 to .77. For item #41 (a sentence revision task) on the same form, an opposite situation occurred. Here, three of the four raters expanded on the key, but did so consistently among themselves. The fourth rater followed the key strictly, generally crediting only those sentences that exactly matched the ones listed in the guide. Removing this rater's scores increased agreement from .29 to .73.

On other items, disagreement was more widely evident, largely because the key failed to provide enough guidance or its guidance was not completely correct. For form B item #6 (sentence completion), the key gave four examples and the direction to credit any "noun that makes semantic sense." This direction apparently left too much room for judgment, with some raters awarding credit for completions like "awareness" and "root" in the phrase "the _____ of her conscience." For item #35 on form C (requiring punctuation of a passage), the key provided only a single example of a correct response when many

correct responses were possible, and when correct and incorrect responses could not be easily distinguished because of the complexity of the passage. Finally, in more than one instance the key listed a finite set of correct responses that turned out not to be exhaustive. On form C item #19, a word rearrangement task, the key indicated as acceptable a set of sentences about computer literacy. Many students, however, constructed sentences like the following: "Before the 1980s a major issue was in literacy not computer education." Some raters apparently believed such sentences, though perhaps awkward, to be correct, whereas others did not.

Scoring disagreement was also evidenced on the multiple-choice items. Some disagreement is expected even for these putatively objective questions because of both the fallibility of humans and the experimental nature of the grading. In most cases, the exact cause of disagreement is difficult to infer as the examinee's response is extremely limited (e.g., a check mark next to an answer option) and because the reason for the rater's grading was not notated. Random errors might be caused by misreading the key, among other things. More consistent inaccuracies might be generated by grading without the key, perhaps using an incorrect memory of it or relying on incomplete content knowledge in place of it. Such systematic inaccuracy should be associated with particular raters. Two of the three multiple-choice items with the lowest generalizability coefficients showed this pattern: removing one of the raters from form C #18 raised the generalizability coefficient from .72

to .90; an equivalent deletion for form A item #26 changed the coefficient from .89 to .95.

In some instances, however, the causes of disagreement in scoring multiple-choice items could be more definitively inferred. This situation was true of form D item #21, which had a generalizability coefficient of .80. This item asked the examinee to identify the error in a sentence by marking the letter that corresponded to the underlined phrase containing the error. Several examinees chose two options--either by indicating two letters or by writing in corrections for two phrases. Credit for these responses was given by some raters if the correct option was included.

In Table 7 the median generalizability coefficients are shown stepped up for scores produced by multiple gradings. For example, the mean of two ratings for selection/identification produces a value equivalent to a single-rating for the multiple-choice category, but four ratings are needed to achieve this level for construction or substitution/correction items.

Insert Table 7 about here

Discussion

This paper presented an initial scheme for classifying constructed response items and the results of two analyses of it. The first analysis found substantial agreement in the assignments of items to the categories; over all items, the median number of judges disagreeing with the intended categorization of an item

was 2 (out of the 27 responding). There was, however, more disagreement in some cases than might have been expected given this consensus; even one of the multiple-choice items was classified differently from its intended assignment by one-fourth of the judges.

A possible explanation for many of the disagreements involves a confusion between two characteristics of an item--what the examinee is expected to accomplish, and how that is to be done. For example, multiple-choice item #13 requires the examinee to fill in a blank with the word that best fits the meaning of a sentence, and to do this by choosing one of five alternatives presented. An apparent focus on "what"--fill in the blank--rather than the intended "how"--by choosing among options --led a number of judges to classify this item as one of completion.

It seems plausible to conjecture that many, if not most, such disagreements could be eliminated by providing more detailed instructions and examples to judges. A few ambiguities would remain and might require elaborating the definitions of some categories. For example, the description of the reordering/ rearrangement category might be extended to state explicitly that assignment to this category requires that the elements presented are to be rearranged with no modification whatsoever and with no addition of further elements, however minor; such an explanation might have forestalled the large number of judgments placing a construction shift item in this category.

So far as differences among the categories are concerned, selection/identification stands out as yielding a higher proportion of items on which there was appreciable disagreement than any other category. All eight items in this category exceeded the median number of disagreements for the entire item set. However, this category was represented by only two item formats, not necessarily a representative sample of those that could be created. (Moreover, three of the six multiple-choice items, included in the analysis to provide a baseline against which to compare other formats, also produced more than the median number of disagreements.) It would be premature to conclude that any one category is inherently less clearly identifiable than the others.

Turning to the study of agreement in scoring students' responses, what is most salient is the high variation across and within categories. Differences across categories, such as those between multiple-choice and reordering/rearrangement, suggest that at least some meaningful distinctions can be identified through this classification scheme. The wide variation <u>within</u> such categories as completion and substitution/correction is somewhat artifactual, due in part to ambiguities in the keys for specific items or the actions of individual raters. Still, even accounting for these correctable factors some variation remains, suggesting the need to subdivide some item categories.

As expected, the study detected a significant relationship between the openness of the response and scoring agreement. This relationship was moderate, with the most open responses--those to

the three construction items--yielding reasonably good agreement. Underlying this relation would appear to be the openness of the scoring key. In the case of the construction items, each had a very detailed guide noting the components that should be included as well as the number of points to be credited or debited for specific features of the writing. Some of the latter required judgment in scoring; for example, determining whether to deduct a point for "formatting the information in an inefficient or disorganized way." Such a requirement was evidently less a source of disagreement than the scoring of more structured items in which it was not possible to provide an exhaustive key and judges had to determine whether an answer was close enough to an ideal response to receive credit.

Several judges participating in this study provided comments critical of the keys to some of the items they were asked to score. Many of these criticisms were well-founded--there were instances in which a purportedly exhaustive key was not exhaustive, as well as other errors and ambiguities in the formulae specified for deriving item scores. It is evident in retrospect that as much effort must be invested in reviewing and revising the keys for such items as is invested in the items themselves. It was not possible, given the limits on time test development staff could devote to an experimental investigation, to subject either the items or the keys to the exhaustive reviews typical for conventional items, much less to the still more demanding reviews we would now expect to be required for items like those employed here. Because of this fact, and because of

the newness of the tasks required of those participating, it is reasonable to view these results as an underestimate of what might be obtained in operational use of these item types.

Of some import in considering these results is whether scoring agreement is good enough--or could be made good enough with experience, a very thorough review process, and perhaps multiple ratings--to permit using these item formats in "high stakes" tests. Some preliminary judgments can be derived using as benchmarks the admittedly imperfect multiple-choice coefficients for an upper bound and those commonly found for essays as a lower one. Comparisons also need to consider the similarity of the item category with these benchmarks in terms of the openness and length of the response. For essay items, reasonable approximations might be the single-reader coefficients reported in Breland et al.'s comprehensive writing assessment investigation, which ran from the low .50s to mid .60s (Breland, Camp, Jones, Morris, & Rock, 1987).

Using these standards, it would seem as if selection/identification items come reasonably close to duplicating "objective" levels of agreement: the median and range for this category are very similar to those for multiple-choice. The construction items, which approximate the length and complexity of essay responses, compare favorably with Breland et al.'s values. The remaining category medians are noticeably lower than the multiple-choice value but at least as good or substantially better than the figures found for essay items.

The usability of these formats therefore appears to depend heavily on whether some degree of error is acceptable, as is the case in evaluating productions such as essays, or whether absolute accuracy is required, as has been typical of conventional "high stakes" tests. This decision also needs to weigh the differential benefits gained from the categories (e.g., the categories differ in response complexity and, thus, have different implications for face validity, instructional diagnosis, and influencing teaching and learning). Finally, the fact should be considered that aggregations over even a small set of items imply a relatively small scoring error overall (but not the elimination of such error altogether).

The present study was an exploratory one, attempting to elicit information about the characteristics of a broad sampling of item formats. One appropriate direction for further work would be to select a limited number of these formats for more rigorous investigation. Items and keys would be developed with the same series of reviews and revisions employed for operational tests, data from a pretest sample would be scored, and items and keys would again be revised. A further data collection and scoring would then provide a more precise indication of the accuracy of scoring likely to be obtainable in practice.

If the preliminary framework presented here seems to provide a useful organizing rubric, it would also be worthwhile to attempt to apply it to sets of items drawn from other content domains. The limitation to verbal items in the present study was deliberate, an attempt to avoid confounding differences in

formats with differences in content; but the scheme is intended to be more general, and its generalizability merits examination.

If the framework does successfully generalize to other contents, a next step might be to examine its empirical validity. Do correlations among items tend to be greater within than across categories? Does analysis of the cognitive demands of the various formats suggest greater similarities within categories?

Finally, it might be useful to consider additional dimensions along which items could be categorized. The one-dimensional scheme presented here considers only the raw material from which the response is selected or constructed. Surface features of the demands of the task, as well as the degree to which an open item can yield a closed key, are among the additional dimensions that might be added to the scheme.

References

Bennett, R. E. (In press). Toward intelligent assessment: An integration of constructed response testing, artificial intelligence, and model-based measurement. In N. Frederiksen, R. J. Mislevy, and I. Bejar (Eds), Test theory for a new generation of tests, Hillsdale, NJ: Lawrence Erlbaum Associates.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). Assessing writing skill. New York: College Entrance Examination Board.

Carlson, S. (1985). Creative classroom testing: Ten designs for assessment and instruction. Princeton, NJ: Educational Testing Service.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley & Sons.

Fiske, E. B. (1990, January 31). But is the child learning? Schools trying new tests. The New York Times, pp. 1, B6.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18(9), 27-32.

Frederiksen, N. (1984). The real test bias. American Psychologist, 39, 193-202.

Jensen, E. (1987). Grading the Advanced Placement Examination in English Language and Composition. New York: College Entrance Examination Board.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Nickerson, R. S. (1989). New directions in educational assessment. Educational Researcher, 18(9), 3-7.

Thorndike, R. L. (1982). Applied psychometrics. Boston, MA: Houghton Mifflin.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. Journal of Educational Measurement, 17, 11-29.

Winer, B. (1971). Statistical principles in experimental design (Second edition). New York: McGraw-Hill.

Table 1

Variance Components for Classification into Six Categories of 36 Items by 27 Judges

Variance Component	Sum of Squares	df	Mean Square	F	P	Variance Estimate
Category	2353.10	5	470.62			2.90
Judge	45.46	26	1.75	4.80	.001	.04
Items within categories	21.68	30	.72	1.98	.01	.01
Category-by-judge	120.51	130	.93	2.54	.001	.09
Item-by-judge within category	284.32	780	.37			.37

Table 2

Frequency Distribution of Product-Moment Correlations Between Judges' Categorizations and the Intended Item Categorizations for 27 Judges

Correlation	Frequency
.65 - .79	2
.80 - .84	2
.85 - .89	3
.90 - .94	9
.95 - 1.00	11

Table 3

Frequency Distribution of Mean Differences Between Each of 27 Judges' Categorizations and the Intended Item Categorizations

Mean Signed Difference	Frequency
-.74 to -.50	0
-.49 to -.25	0
-.24 to -.01	4
0 to .24	19
.25 to .49	2
.50 to .74	2

Table 4

Number of Judges Whose Item Categorizations Diverged from the Intended Item Categorizations

Item Number	Category	Item Descriptor	# of Judges Diverging
36	Multiple-choice	sentence identification	1
18	Multiple-choice	error location	2
26	Multiple-choice	sentence identification	2
1	Multiple-choice	error location	3
21	Multiple-choice	error location	3
13	Multiple-choice	sentence completion	7
38	Selection/Identification	cloze elide	3
2	Selection/Identification	cloze elide	4
37	Selection/Identification	cloze elide	4
32	Selection/Identification	keylist	5
4	Selection/Identification	keylist	7
9	Selection/Identification	cloze elide	7
10	Selection/Identification	cloze elide	7
43	Selection/Identification	keylist	8
12	Reordering/rearrangement	sentence ordering	0
19	Reordering/rearrangement	word rearrangement	0
28	Reordering/rearrangement	word rearrangement	0
44	Reordering/rearrangement	sentence ordering	0
40	Reordering/rearrangement	word rearrangement	1
16	Reordering/rearrangement	sentence ordering	1
30	Reordering/rearrangement	word ordering	7
24	Reordering/rearrangement	classification	8
25	Substitution/correction	word correction	1
45	Substitution/correction	word substitution	2
8	Substitution/correction	word substitution	2
29	Substitution/correction	word substitution	2
34	Substitution/correction	word substitution	3
31	Substitution/correction	sentence combining	6
46	Substitution/correction	sentence combining	6
5	Substitution/correction	sentence combining	9
41	Substitution/correction	construction shift	17
33	Completion	word cloze	0
39	Completion	word cloze	0
6	Completion	sentence completion	0
22	Completion	sentence completion	1
27	Completion	word insertion	3
17	Completion	paragraph completion	4
3	Completion	word insertion	10
35	Completion	punctuation	12
7	Construction	letter writing	0
11	Construction	essay	0
15	Construction	essay	0
20	Construction	announcement writing	0
23	Construction	essay	0
42	Construction	essay	0
14	Construction	short explanation	1

Table 5

Generalizability Coefficients for a Single Rater's Item Scores

Item Number	Category	Item Descriptor	Generalizability Coefficient
		Form A	
1	Multiple-choice	error location	1.00
26	Multiple-choice	sentence ident.	.89
2	Selection/identification	cloze elide	.92
32	Selection/identification	keylist	.87
5	Substitution/correction	sentence combining	.83
31	Substitution/correction	sentence combining	.75*
12	Reordering/rearrangement	sentence ordering	.87
28	Reordering/rearrangement	word rearrangement	.63*
27	Completion	word insertion	.93
7	Construction	letter writing	.78*
		Form B	
13	Multiple-choice	sentence completion	.97
36	Multiple-choice	sentence ident.	.98
4	Selection/identification	keylist	.85
16	Reordering/rearrangement	sentence ordering	.53*
30	Reordering/rearrangement	word ordering	.49*
8	Substitution/correction	word substitution	.98
45	Substitution/correction	word substitution	.95
34	Substitution/correction	word substitution	.60*
33	Completion	word cloze	1.00
6	Completion	sentence completion	.23*
14	Construction	short explanation	.84
		Form C	
18	Multiple-choice	error location	.72
43	Selection/identification	keylist	.94
9	Selection/identification	cloze elide	.72
19	Reordering/rearrangement	word rearrangement	.34*
25	Substitution/correction	word correction	.74
46	Substitution/correction	sentence combining	.30*
17	Completion	paragraph comp	.71
39	Completion	word cloze	.63
35	Completion	punctuation	.41*
20	Construction	announcement writing	.57
		Form D	
21	Multiple-choice	error location	.80
10	Selection/identification	cloze elide	.90
38	Selection/identification	cloze elide	.86
37	Selection/identification	cloze elide	.77
44	Reordering/rearrangement	sentence ordering	.81
24	Reordering/rearrangement	classification	.59*
40	Reordering/rearrangement	word rearrangement	.52*
29	Substitution/correction	word substitution	.71
41	Substitution/correction	construction shift	.29*
3	Completion	word insertion	.73
22	Completion	sentence completion	.42*

*Agreement coefficient is in bottom third for the test form.

Table 6

Median and Range of Generalizability Coefficients for a Single Rater's Scores with Items Collapsed Across Test Forms within Categories

Item Category	Median Coefficient	Range
Multiple-choice (6)	.93	.72-1.00
Selection/identification (8)	.87	.72- .94
Reordering/rearrangement (8)	.56	.34- .87
Substitution/correction (9)	.74	.29- .98
Completion (8)	.67	.23-1.00
Construction (3)	.78	.57- .84

Note. The number of items in a category is shown in parentheses.

Table 7

Median Generalizability Coefficients for Item Scores for Different Numbers of Raters

Item Category	Number of Raters			
	One	Two	Three	Four
Multiple-choice (6)	.93	.97	.98	.98
Selection/identification (8)	.87	.93	.95	.96
Reordering/rearrangement (8)	.56	.73	.79	.84
Substitution/correction (9)	.74	.85	.89	.92
Completion (8)	.67	.80	.86	.89
Construction (3)	.78	.87	.91	.93

Note. The number of items in a category is shown in parentheses.

Figure 1

A Scheme for Categorizing Item Types

---

0. Multiple-choice: Items in this class require the examinee to choose an answer from a small set of response options.

Example. Choose the word which, when inserted in the sentence, best fits the meaning of the sentence as a whole.

Unable to focus on specific points, he could talk only about _____; indeed, his entire lecture was built around vague ideas.

(A) personalities
(B) statistics
(C) vulgarities
(D) particulars
(E) abstractions

1. Selection/Identification: This category is characterized by choosing one or more responses from a stimulus array. In contrast to multiple-choice, the number of possible choices is typically large enough to limit drastically the chances of guessing the correct answer. In addition, in its ideal form, the response to this item type is probably mentally constructed and not simply recognized. Examples include keylists, cloze elide (i.e., deleting extraneous text from a paragraph), and, via touch screen, tracing orally presented directions on a computer generated map.

Example. Delete the unnecessary or redundant words from the following paragraph:

Andy Razaf is not a quickly recognizable name that is familiar to most people. Yet Razaf wrote the lyrics to at least 500 or more songs, including the words to the popular "Ain't Misbehaving'," "Honeysuckle Rose," and "Stompin' at the Savoy" as well. The American-born son of an upper class African nobleman, he still continues to be overshadowed by his composer-collaborators who worked with him, Fats Waller and Eubie Blake.

Figure 1 (con't)

A Scheme for Categorizing Item Types

---

2. Reordering/rearrangement: Here, too, responses are chosen from a stimulus array. However, the task in this case is to place items in a correct sequence or alternative correct sequence. Examples include constructing anagrams, ordering a list of sentences to make them reflect a logical sequence, categorizing elements in a list, arranging a series of mathematical expressions to form a correct proof, arranging a series of pictures in sequence, and putting together a puzzle.

> Example. Rearrange the following group of words into a complete and meaningful sentence. Capitalize the first word and end with a period. No other marks of punctuation should be needed.
>
> a and be both can comedy enlightening entertaining good

3. Substitution/correction: This item type requires the examinee to replace (as opposed to reorder or rearrange) what is presented with a correct alternative. Examples include correcting misspellings, correcting grammatical errors, substituting more appropriate words in a sentence, replacing several sentences with a single one that combines the meanings of each, correcting faulty computer programs, and substituting operators to create a true mathematical expression.

> Example. Combine the two sentences below into one grammatically correct sentence that conveys the same information as the original pair.
>
> 1. Stephen King is the author of numerous horror novels.
>
> 2. Many fans of Stephen King assume that he is as crazy as some of his characters.

4. Completion: In this item type, the task is to respond correctly to an incomplete stimulus. Cloze, sentence completion, mathematical problems requiring a single numerical response, progressive matrices, and items that require adding a data point to a graph when given appropriate numerical data are examples.

> Example. Fill the blank in the following sentence with one word that makes the sentence grammatically and logically complete.
>
> Melodramas, _____ present stark contrasts between good and evil, are popular forms of entertainment because they offer audiences a world where there is moral certainty.

Figure 1 (con't)

A Scheme for Categorizing Item Types

---

5. Construction: Whereas the Completion type requires that a stimulus be completed, here construction of a total unit is required. Examples are drawing a complete graph from given data, listing a country's exports, stating why condensation forms on windows, writing a geometric proof, producing an architectural drawing, and writing a computer program or essay.

> Example. Describe some event or phenomenon in the natural world (e.g. earthquakes, thunderstorms, rainbows) that has always interested you and that you would like to know more about. What in particular would you like to know about this subject, and why? (You will have 1/2 hour in which to write this essay.)

6. Presentation/Performance: This item type requires a physical presentation or performance delivered under real or simulated conditions in which the object of assessment is in some substantial part the manner of performance and not simply its result. Examples include repairing part of an automobile engine, playing an instrument, diagnosing a patient's illness, teaching a demonstration lesson, giving a theatrical audition.

> Example. Perform two contrasting solo pieces not to exceed two minutes each. Timing begins with an introduction in which you announce the audition in the following manner: "My name is (give name). My first piece is from (title of play) by (author). I play the part of (character). My second piece is from (title of play) by (author). I play the part of (character)." Props are limited to one stool, two chairs, and one table. To allow you to show your versatility, it is to your advantage to have the greatest possible contrast between your pieces. You will be judged on your ability to demonstrate control of material; flexibility of voice, movement, and expression; and vocal and physical articulation.

Appendix A: Items Organized by Category

0. Multiple Choice

1. The following sentence may contain an error in one of the underlined portions. If so, indicate below the letter of the portion that contains the error. If the sentence is correct as written, mark "E."

Once (A) Art Deco is called to your attention, one sees (B) its influence everywhere, (C) in theater lobbies, in furniture design, even in (D) perfume bottles. No error (E)

A. _____
B. _____
C. _____
D. _____
E. _____

13. Choose the word which, when inserted in the sentence, best fits the meaning of the sentence as a whole. Unable to focus on specific points, he could talk only about _____; indeed, his entire lecture was built around vague ideas.

(A) personalities
(B) statistics
(C) vulgarities
(D) particulars
(E) abstractions

18. The following sentence may contain an error in one of the underlined portions. If so, circle the letter of the option that contains the error. If the sentence is correct as written, mark "E."

With the invention of (A) the hypodermic syringe and the administration of pure morphine in large numbers to (B) wounded soldiers during the Civil War, narcotics addiction became (C) a serious social problem (D) in the United States.

No error (E)

A. _____
B. _____
C. _____
D. _____
E. _____

21. The following sentence may contain an error in one of the underlined portions. If so, cirle the letter of the option that contains the error. If the sentence is correct as written, mark "E."

As much (A) as 200 North American Indian (B) languages and dislects have ceased (C) to exist in that there are no (D) surviving speakers or written records.

No error (E)

A. _____
B. _____
C. _____
D. _____
E. _____

26. Indicate which of the following sentences is grammatically correct and best expresses its meaning.

(A) Mass determines whether a star will compress itself into a "white dwarf," a "neutron star," or a "black hole" after it passes through the "red giant" stage of its life cycle.

(B) A star's compression of itself will be a "white dwarf," a "neutron star," or a "black hole" after it passes through the "red giant" stage of its life cycle, depending on its mass.

(C) After passing through the "red giant" stage of its life cycle, depending on a star's mass, a star will compress until there is a "white dwarf," a "neutron star," or a "black hole."

(D) After passing through a "red giant" stage of a life cycle, a star's mass will determine if the compression of itself is into a "white dwarf," a "neutron star," or a "black hole."

(E) The mass of a star, after passing through a "red giant" stage of a life cycle, will determine whether or not to compress itself into a "white dwarf," a "neutron star," or a "black hole."

36. Indicate which of the following sentences is grammatically correct and best expresses its meaning.

(A) Licht did not realize he was being filmed, and when he was caught by the movie camera, he was eating a fish that still had its head on and was drinking red wine in great gulps.

(B) Licht did not realize he was being filmed, and when he was caught by the movie camera, he was eating a fish with its head still on, drinking red wine in great gulps.

(C) Licht did not realize he was being filmed, and when he was caught by the movie camera, he had been eating a fish with its head still on and was drinking red wine in great gulps.

(D) Licht did not realize he was being filmed, and when he was caught by the movie camera, he had been drinking red wine in great gulps as he is eating a fish that still had its head on.

(E) Licht did not realize he was being filmed, and when he was caught by the movie camera, he was drinking red wine in great gulps and eating a fish with its head still on.

1. Selection/Identification

2. The following passage contains irrelevant or incorrect words that interfere with the meaning or produce grammatical errors. Delete these words so that the writing is grammatical and the sense of the passage is not disrupted.

Ludwig van Beethoven's life was not specific particularly rich in external events: great occasions were rare puzzles, and he never traveled to otherwise distant places. He spent almost all his life in the cities of Bonn and Vienna, working on his music. Unlike that of Mozart, who had seen much of Europe during his concert tours while still a boy, Beethoven went on very few journeys after regrets moving to Vienna ordinarily in November of 1792, at the age of 21. A concert tour ago to Prague and Berlin in 1796 and another to Prague in 1798 were exceptions; in general, he vastly left Vienna and its immediate surroundings only occasionally, and when he did it afterwards was to spend a week or so as the guest of aristocratic patrons and friends.

When Beethoven arrived in Vienna he was still despite a member of the Bonn court orchestra, as he had been ever since the age of 14, but with the extra collapse of the government at Bonn a few years later he was left entirely to his own devices. Instead of being able while to enjoy the security of a court musician's post, as his father and grandfather had then before him, he was forced to find ways to earn his uncertainty living purely through his work as a composer, virtuoso pianist, and conductor. These problems were equal followed by another far more serious: gradually increasing when deafness, which finally deprived him of help the ability to hear his own music performed. To this severe trial was added the death of his revoked brother Karl in 1815. Thereafter, Beethoven assumed never guardianship of Karl's profligate and dissolute son, a responsibility that caused Beethoven ending much personal as well as financial embarrassment. The effect of all these tribulations can be seen clearly in Waldmuller's scholarship portrait of the aging master.

4. The following passage contains underlined portions that represent possibly inappropriate word choice. Read the entire passage. Then, for each underlined word that represents an inappropriate word choice, think of a more appropriate choice and look for it on the list below. Write the word from the list just above the underlined word in the passage.

A brainy naturalist once stated that among the many riddles of nature, not the least arcane is the migration of fishes. The homing of salmon is a particularly bold example. The Chinook salmon of the U.S. Northwest is born in a small stream, migrates downriver to the Pacific Ocean as a young smolt and, after living in the sea for as long as five years, swims back infallibly to the stream of its birth to procreate. Its determination to return to its birthplace is mythical. No one who has seen a 100-pound Chinook fling itself into the air again in a useless effort to overcome a waterfall can fail to marvel at the strength of the instinct that draws the salmon upriver to the place where it was born.

NO CHANGE REQUIRED	economic	reliable
appreciable	experiment	sails
astronomical	fascinated	spawn
belittle	fictitious	surmount
biased	frugal	theatrical
centuries	heavily	tremendous
conceited	immature	defeat
condescending	invisible	understood
conjugate	learned	unerringly
cybernetic	legendary	unstintingly
deeply	luring	vain
defeat	molt	vane
designated	monumental	vault
designed	mysterious	violent
despair	nesting	whim
destination	perigrinates	NO APPROPRIATE REPLACEMENT LISTED
different	purely	
dramatic	rarely	
	regimented	

9. The following passage contains irrelevant or incorrect words that interfere with meaning or produce grammatical errors. Delete these words so that the writing is grammatical and the sense of the passage is not disrupted.

Just such enough is known about Phillis Wheatley's life to suggest the able extent of her poetic talent, for she heard developed it against great odds. The time and place of Wheatley's similar birth are as unknown as these of her African name, but she probably came from whether what is now called Senegal or Gambia. Purchased directly off limits a slave ship in Boston by a wealthy tailor, John Wheatley, in primarily 1761, she was losing her first teeth, and so she was believed to be rich about seven years of age. She learned English in sixteen months, and soon more studied Latin as well as the Bible and English poetry by Alexander Pope and Thomas Gray. She began writing sudden religious verse when she was then thirteen, and she could not have still been more than seventeen years old when she published her first poem, announced an elegy on the death of the English evangelical preacher George Whitehead.

10. The following passage contains irrelevant or incorrect words that interfere with the meaning or produce grammatical errors. Delete these words by crossing them out so that the writing is grammatical and the sense of the passage is not disrupted.

It's worth the drive trip to Medford to enjoy the valley's best and finest Mexican-American restaurant place, "Mexican Rose." Every day daily specials of fresh new charbroiled seafood, traditional dishes, steaks and ribs, and even also vegetarian good meals are served in an art deco atmosphere. Try one of their exotic drink libations at the bar, or a pitcher of margueritas with dinner. "Mexican Rose" was voted the best top Mexican restaurant in the region area.

32. The word that best completes the sentence below appears in the alphabetical word list that follows the sentence. Put the number of this word in the blank space.

The gravitational force of a "black hole" in space is _____ strong that not even light can escape it: any beam that enters the field gets pulled into the so-called hole, where it remains trapped.

1.	actually	11.	distant	21.	never	31.	so
2.	afterwards	12.	enough	22.	nevertheless	32.	such
3.	also	13.	especially	23.	no	33.	that
4.	although	14.	extremely	24.	not	34.	therefore
5.	as	15.	force	25.	notably	35.	this
6.	awfully	16.	how	26.	otherwise	36.	too
7.	because	17.	however	27.	overly	37.	unknown
8.	consequently	18.	like	28.	probably	38.	very
9.	despite	19.	more	29.	really	39.	whether
10.	discovered	20.	most	30.	since	40.	while

37. Delete the unnecessary or redundant words from the following paragraph:

Andy Razaf is not a quickly recognizable name that is familiar to most people. Yet Razaf wrote the lyrics to at least 500 or more songs, including the words to the popular "Ain't Misbehavin'," "Honeysuckle Rose," and "Stompin'at the Savoy" as well. The American-born son of an upper-class African nobleman, he still continues to be overshadowed by his composer-collaborators who worked with him, Fats Waller and Eubie Blake.

38. The following passage contains irrelevant or incorrect words that interfere with meaning or produce grammatical errors. Delete these words so that the writing is grammatical and the sense of the passage is not disrupted.

"Dickens," George Orwell once remarked, "is one of those writers well worth imitating." Consequently, many different fraction groups were eager to claim him as were one of their own comatose. Did Orwell foresee as that someday he too would become just nicely such as a writer? Almost certainly incomplete he did not. In 1939, when he wrote dogged those words about Dickens, Orwell was still a true relatively obscure figure and among dishes those who knew his work at all wrongs, a highly controversial finally-one. Only a year earlier, than his work had been extent rejected on political grounds flag by his own publishers in both Britain and the United States tomorrow. Nevertheless and, by the time hearing of his death in 1950 at the age slightly of forty-six, he had become old so famous today that his very name entered regret the lnguage and has remained tight there in the form of the adjective "Orwellian" birds.

43. The word that best completes the sentence below appears in the alphabetical word list that follows the sentence. Put the number of this word in the blank space.

Even when they are isolated from sunlight, plants are still able to tell _____ it is day or night.

1.	actually	11.	distant	21.	never	31.	so
2.	afterwards	12.	enough	22.	nevertheless	32.	such
3.	also	13.	especially	23.	no	33.	that
4.	although	14.	extremely	24.	not	34.	therefore
5.	as	15.	force	25.	notably	35.	this
6.	awfully	16.	how	26.	otherwise	36.	too
7.	because	17.	however	27.	overly	37	unknown
8.	consequently	18.	like	28.	probably	38.	very
9.	despite	19.	more	29.	really	39.	whether
10.	discovered	20.	most	30.	since	40.	while

BEST COPY AVAILABLE

2. Reordering/Rearrangement

12. The four sentences in the following paragraph are out of order. Logically reorder them by indicating in parentheses what number each sentence should have been in the revised paragraph.

( ) However, if the star was originally more massive, equal to three or four of our Suns, it compresses further and changes from a "white dwarf" into a "neutron star." ( ) At the end of its life cycle, a star begins to compress after it has burned up all of its hydrogen and helium. ( ) And if the original star was still more massive, the neutron star continues compressing until it crushes itself into that most mysterious of all forms in outer space, a "black hole." ( ) If the star was originally less massive than about two of our Suns, it compresses until it becomes a "white dwarf."

16. The four sentences in the following paragraph by Alfred Hitchcock are out of order. Logically reorder them by indicating in parentheses what number each sentence should have in the revised paragraph.

( ) Unfortunately, few of the books seemed to have much connection with what one saw at the local movie theater. ( ) Nobody wrote for the sensible middlebrow moviegoer who was keenly interested in the craft of the cinema without wanting to make a religion of it. ( ) Thirty or forty years ago, when the idea of the cinema as an art form was new, people started to write highbrow treatises about it. ( ) Even earlier began the still-continuing deluge of fan magazines and annuals, full of exotic photographs but short on solid information.

19. Rearrange the following group of words into a complete and meaningful sentence. Capitalize the first word and end with a period. No other marks of punctuation should be needed.

a the in not was 1980s issue literacy major before education computer

24. The following is an alphabetical list of subjects people study at universities. Re-order and classify these subjects into four or five categories that represent major fields or disciplines. Label your categories and give a brief explanation of your system of classification.

Accounting	Foreign Languages
Anatomy	Forestry
Anthropology	History
Archaeology	Law
Architecture	Linguistics
Biology	Marine Biology
Business	Mathematics
Chemistry	Mechanical Engineering
Chemical Engineering	Music
Computer Science	Neurology
Dance	Philosophy
Drama	Physics
Earth Science	Political Science
Economics	Psychology
Education	Sociology
English	Urban Studies
Finance	Women's Studies
Fine Arts	Zoology

28. Rearrange the following group of words into a complete sentence. Capitalize the first word and end with a period. No other marks of punctuation should be needed.

a and be both can comedy enlightening entertaining good

30. Make as many grammatically correct English sentences as you can using only words from the following list. A sentence may use any number of words from the list, but a word can appear only once in any sentence.

an
extremely
fish
have
of
sense
sensitive
smell

40. Rearrange the following group of words into a complete sentence. Capitalize the first word and end with a period. You may add punctuation if you feel it is needed.

fewer age to people as tend they colds get

44. The five sentences in the following paragraph are out of order. Logically reorder them by indicating in the parentheses what number each sentence should have in the revised paragraph.

( ) With the 1986 Tax Reform Act, however, the game plan has changed. ( ) In either case, the years a person has already worked for his or her present employer count. ( ) How does one get vested in a company pension plan these days? ( ) Scheduled to take effect this year, the new rules reduce the vesting period to five years--to partial vesting after three years with full vesting after seven. ( ) Until this year most workers had to be employed by a company for ten years before they became vested--that is, entitled to received a pension at retirement.

3. Substitution/Correction

5. Combine the two sentences in (A) by writing a phrase in the blank in (B) that makes (B) a single grammatical sentence. This sentence should contain the same information and have the same meaning as the pair in (A).

(A) The discovery of "black holes" is among the most exciting recent developments in astronomy. It came well after the discovery of "red giant" stars.

(B) The discovery of "black holes,"_____, is among the most exciting recent developments in astronomy.

8. Correct the following sentence by crossing out the one word that produces a grammatical error and substituting the appropriate word.

Many fans of Stephen King, the author of numerous popular horror novels,

assume that he is so mad as some of his characters.

25. Cross out the words in the passage below that are misspelled. Write the word correctly in the space at the right of the lines. If there are no misspellings in the line, write nothing on the line.

Sometimes pruning is called a sceince.	_____
Sometimes pruning is called an art. The	_____
defenition depends on the purpose. For the	_____
average gardner, pruning is a means of keeping	_____
plants under control to fill their alloted	_____
spaces. When the plans outgrow their spaces,	_____
they must be disaplined.	_____
Either approach requires some knowledge.	_____
Merely hacking with a saw and pruning shears	_____
is not helpfull to the plant's form or	_____
vigor. This is especially true when amature	_____
prumers shape plants from the top only and	_____
fail to get underneath and cut out older	_____
growth. From Febuary until genuine spring when	_____
the buds begin to break, plants are dormant	_____
and can be pruned. This is the time to do some	_____
serious homework and look at some of the	_____
source books on pruning.	_____

29. Correct the following sentence by crossing out the one word that produces a grammatical error and substituting the appropriate word.

The sixteenth-century art critic Vasari regarded the painting entitled the Mona Lisa is a wonderfully faithful reproduction of an actual person; to many nineteenth-century critics, it was a symbol to be decoded.

31. Combine the two sentences below into one grammatically correct sentence that conveys the same information as the original pair.

1. The fires set to fumigate the houses of the victims of the Black Death destroyed many documents.
2. These could have identified the victims and their ancestors.

34. Replace each underlined word or phrase in the passage below with a different word or phrase that changes the meaning of the original as little as possible.

Some faculty members took me and Novella out to lunch in San Jose's finest <u>eatery</u> -- nerves much <u>assuaged</u> by their <u>kindness</u> and several drinks. In midlunch two men came over to our table, a dean and a <u>spruce</u> young fellow looking something like a <u>composite</u> of the junior Watergate <u>get</u> we'd seen on television, who introduced himself as a lawyer for the university trustees. I said, Oh good, I need a lawyer: I just got this <u>absurd</u> note about a loyalty oath and fingerprinting. There's not a <u>word about</u> either in my contract.

41. "The rocky outcrops of North America are still roamed by the bobcat, though it is seldom seen or heard."

Rewrite the sentence above so that it conveys the same meaning as the original. START your new sentence with "The bobcat."

45. Correct the following sentence by crossing out the one word that produces a grammatical error and substituting the appropriate word.

The roads and means of transportation remain as they did thirty years ago; only the town hall with its television aerial is new.

46. Combine the two sentences below into one grammatically correct sentence that conveys the same information as the original pair.

1. Stephen King is the author of numerous horror novels.

2. Many fans of Stephen King assume that he is as crazy as some of his characters

4. Completion

3. Insert words into the sentence below that will make the statement logically and grammatically complete.

Birds, bees, and various migratory species can tell direction they are traveling; for example, a migrating flock can use the positions of the Sun or stars find north.

6. Fill in the blank in the following sentence with a word that makes the sentence grammatically and logically complete.

Anti-apartheid writer Janet Levine attributes the _____ of her conscience to several mentors, not the least of these being a Black family maid who spoke bitterly of the injustices in South Africa.

17. Underneath the paragraph below, write a sentence that could supply the logical connection that is missing from the paragraph. Base your sentence on what has preceded and what follows the space for the sentence.

Archaeologists believe that they have found the site of the Rose Theater, a celebrated sixteenth-century, open-air playhouse where works by Shakespeare, Marlowe, Jonson and other leading Elizabethan playwrights were performed. Since December, a team of twelve archaeologists has studied the site, which was exposed after a 30-year-old office building was razed to make way for a new structure.

_____ But in recent weeks scholars, theater buffs and actors have protested plans to end the dig, written letters to the newspapers, and attempted to negotiate with the property owners.

22. Fill in the blank in the following sentence with a word that makes the sentence grammatically and logically complete.

By lobbying for changes in hunting laws and releasing animals born in captivity into the wild, conservationists are _____ to save or re-establish populations of animals, such as grizzly bears and panthers, that have been systematically trapped, shot, or poisoned nearly to extinction.

27. Insert the word into the sentence below that will make the statement logically and grammatically correct.

The human mind delights finding patterns--so much so that we often mistake coincidence for profound meaning.

33. Fill the blank in the following sentence with one word that makes the sentence grammatically and logically complete.

Melodramas, _____ present stark contrasts between good and evil, are popular forms of entertainment because they offer audiences a world where there is moral certainty.

35. Insert whatever punctuation is needed to make the sentence given below clear and grammatically correct.

This entire allegory I said you may now append dear Glaucon to the previous argument the prison-house is the world of sight the light of the fire is the sun and you will not misapprehend me if you interpret the journey upwards to be the ascent of the soul into the intellectual world according to my poor belief which at your desire I have expressed rightly or wrongly God knows

39. Fill in the blank in the following sentence with one word that makes the sentence grammatically and logically complete.

Kate Millett's Sexual Politics (1970) has been regarded as one of the most important texts of the modern feminist movement, _____ its author is renowned as one of the movement's founders.

5. Construction

7. _____

Golden News April 20, 1977

SUMMER EMPLOYMENT

EARN & LEARN

Positions opening soon for apprentices in

Medical Services

Food Services

Library Services

Earn $3.50 or more per hour while you

learn a valuable skill.

Send letter of application to:

Tyland Training Center
Box 335
Tyland, CA 99499

Pretend that you are Pat Carson and live at 291 Westover Street in Tyland, California. Write a letter applying for the work-training program in one of the catagories listed in the advertisement. You may either give facts about yourself or make up information that you think will help you be accepted.

11. Directions: Please write an essay on ONE of the following topics. (You will have 45 minutes in which to write this essay.)

1. "Ours is an age of indifference--a time when people show little interest in social and political issues."

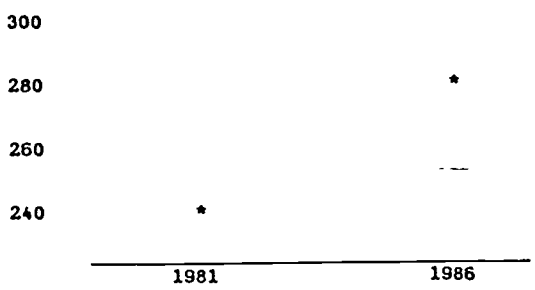
Do you agree or disagree with this statement? In your essay, provide examples to support your view.

2. "Everything in life changes."

Identify one thing in society that has changed significantly in this century. Explain how this change has affected our lives. Be specific.

14. You are preparing a report on endangered species of animals. Write one or two sentences in which you present as much of the information provided by the following graph as possible.

Gorilla Sightings in the Virunga Mountains
of Rwanda, Zaire, and Uganda



Census published in 1986

15. Describe some event or phenomenon in the natural world (e.g. earthquakes, thunderstorms, rainbows) that has always interested you and that you would like to know more about. What in particular would you like to know about this subject, and why? (You will have a 1/2 hour in which to write this essay.)

BEST COPY AVAILABLE

20. You are going to read a transcript of a telephone conversation between two people. After you have read the conversation, write the announcement that you think Pat Carson should put on the bulletin board.

Conversation Transcript

Mrs. Stone: Hello. Pat. This is Vera Stone.

Pat Carson: I thought you were away.

Mrs. Stone: Not until tomorrow. But did you read the newspaper this morning? About the Youth Center?

Pat Carson: No, what happened?

Mrs. Stone: The wind storm did a lot of damage to the roof and grounds. The Youth Center staff will need a a lot of help to get it back in shape.

Pat Carson: I'll be glad to help.

Mrs. Stone: Great, but we'll need a crew of workers. See if you can get about 20 volunteers. Could you put up an announcement outside the principal's office?

Pat Carson: Sure, I'll be glad to.

Mrs. Stone: I'd like to meet on Saturday morning, but I think a lot of the kids have band practice, so let's meet at 1:00.

Pat Carson: That's been cancelled. Why not have them come at 8:30?

Mrs. Stone: Fine. They should bring tools.

Pat Carson: Like what?

Mrs. Stone: Hammers, rakes, shovels -- wheelbarrows if they can. They shouldn't bring any power tools, though. That's all we need, an accident with a power tool. They can work til noon and I'll provide lunch for everybody.

Pat Carson: Great. Then they'll be sure to come. Oh, by the way, do you mean this Saturday or the next?

Mrs. Stone: This one, March 21st.

Pat Carson: Sure, Mrs. Stone. I'll be glad to put up an announcement.

Mrs. Stone: Thanks, Pat. I appreciate your help.

23. Describe your favorite book, poem, film, or piece of music, explaining what features of the work you find most successful or appealing and what, if anything, could be done to improve it. (You will have a 1/2 hour in which to write this essay.)

42. Which of your possessions would be the most difficult for you to give up or lose? Discuss why. (You will have 30 minutes in which to write this essay.)

Appendix B: Scoring Guide Organized by Item Category

0. Multiple Choice

Unless otherwise specified, items should be scored as "1" (correct) or "0" (incorrect).

1. Answer = B

13. Answer = E - Abstractions

18. Answer = B

21. Answer = A

26. Answer = A

36. Answer = E

1. Selection/Identification

Unless otherwise specified, items should be scored as "1" (correct) or "0" (incorrect).

2. See the attached template for an aid in scoring this item.

   Irrelevant words:

specific	vastly	equal
puzzles	afterwards	when
otherwise	despite	help
that of	extra	revoked
regrets	while	never
ordinarily	then	ending
ago	uncertainty	scholarship

   Responses are scored on a 0-7 scale by subtracting the number of erroneous responses from 21, dividing this figure by three, and rounding to the nearest whole number. (Award a 0 if the result is negative.) An erroneous response is either the failure to delete an irrelevant word or phrase (e.g., "that of"), or the deletion of a word or phrase that belongs in the passage.

   Example: 4 failures to delete
   3 inappropriate deletions
   7 total errors

$$\begin{array}{r} 21 \\ -7 \\ \hline 14/3 \end{array} = 4^{2/}{}_{3} = 5 \text{ (total score)}$$

4. See Scoring template.

   Key:

1. learned, fascinated	7. unerringly
2. No change	8. spawn
3. mysterious	9. legendary
4. No change	10. vain, tremendous
5. dramatic, legendary	11. surmount
6. No change	12. No change

   Treat the words NO CHANGE REQUIRED as equivalent to the absence of an insertion over an underlined word. Score as incorrect the use of NO APPROPRIATE REPLACEMENT when no change is required.

   Responses are scored on a 0-4 scale by dividing the number of correct answers by 3 and rounding to the nearest whole number. (Award a 0 if the result is negative.)

9. Irrelevant words:

such	primarily
able	rich
heard	more
similar	sudden
these	then
of	still
whether	announced
limits	

Responses are scored on a 6-point scale (including 0) by subtracting the number of erroneous responses from 15, dividing this figure by 3, and rounding to the nearest whole number. (Award a zero if the result is negative.) An erroneous response is either the failure to delete an irrelevant word or phrase, or the deletion of a word or phrase that belongs in the passage.

Example: 3 failures to delete
1 inappropriate deletion
4 total errors

$$\begin{array}{r} 15 \\ -4 \\ \hline 11/3 = 3^{2/}{}_{3} = 4 \end{array}$$

10. Key:

Line
1. Drop "drive" or "trip"; "best and" or "and finest"
2. Drop "place"; "Every day" or "daily"
3. Drop "new"
4. Drop "even" or "also"; "good"
5. Drop "drink"
6. Drop "best" or "top"
7. Drop "region" or "area"

Responses are scored on a 0-3 scale by subtracting the number of erroneous responses from 10, dividing this figure by three, and rounding to the nearest whole number. (Award a 0 if the resulting score is negative.) An erroneous response is either the failure to delete an irrelevant word or phrase (e.g., "best and"), or the deletion of a word or phrase that belongs in the passage.

Example: 3 failures to delete
3 inappropriate deletions
6 total errors

$$\begin{array}{r} 10 \\ -6 \\ \hline 4/3 = 1^{1/}{}_{3} = 1 \text{ (total score)} \end{array}$$

32. Key "so"

37. Key:

Line
1. Drop "quickly recognizable"
2. Drop either "at least" or "or more"
3. Drop either "the words to" or "words to the"
4. Drop "as well" and "upper class"
5. Drop "still"
6. Drop "who worked with him"

Responses are scored on a 0-2 scale by subtracting the number of erroneous responses from 7, dividing this figure by three, and rounding to the nearest whole number. (Award a 0 if the resulting score is negative.) An erroneous response is either the failure to delete an irrelevant word or phrase (e.g., "who worked with him"), or the deletion of a word or phrase that belongs in the passage.

Example: $\begin{array}{r} 7 \\ -6 \\ \hline 1/3 = 0 \text{ (total score)} \end{array}$

38. See Scoring template.

Key:

Line
2. fraction
3. were; comatose
4. as; nicely; as
5. incomplete
6. dogged; true
7. dishes; wrongs
8. finally; than
9. extent; flag
10. tomorrow; and
11. hearing; slightly
12. old; today; regret
13. tight
14. birds

Responses are scored on a 0-8 scale by subtracting the number of erroneous responses from 23 dividing this figure by three, and rounding to the nearest whole number. (Award a 0 if the resulting score is negative.) An erroneous response is either the failure to delete an irrelevant word or phrase or the deletion of a word or phrase that belongs in the passage.

Example: 23
-6
17/3 = 6 (total score)

43. Key: "whether"

2. Reordering/Rearrangement

Unless otherwise specified, items should be scored as "1" (correct) or "0" (incorrect).

12. Key: (3), (1), (4), (2), where (3) indicates that the first sentence belongs in the third position. Score on a 0-4 scale by awarding 1 point for each correct placement of a sentence.

Example: (4), (1), (3), (2)
1 point + 1 point = 2 points

For imperfect responses only, in addition to awarding points for absolute placement, grant 1/2 point for each correct sequence of two sentences. For example, the sequence (1), (4), (2), (3) would receive 1 point for sequence: 1/2 point for 4 and 2, and 1/2 point for 2 and 3. Round all scores up to the nearest whole number.

16. Key: (2), (4), (1), (3), where (2) indicates that the first sentence belongs in the second position. Score on a 0-4 scale by awarding 1 point for each correct placement of a sentence.

Example: (3), (4), (1), (2)
1 point + 1 point = 2 points

For imperfect responses only, in addition to awarding points for absolute placement, grant 1/2 point for each correct sequence of two sentences. For example, the sequence (4), (1), (3), (2) would receive 1 point for sequence: 1/2 point for 4 and 1, and 1/2 point for 1 and 3. Round all scores up to the nearest whole number.

19. Key:

* Before the 1980s computer literacy was not a major issue in education.
* Computer literacy was not a major issue in education before the 1980s.
* Computer literacy was not before the 1980s a major issue in education.
* Computer literacy was not in education before the 1980s a major issue.
* In education before the 1980s computer literacy was not a major issue.
* In education computer literacy was not a major issue before the 1980s.
* Not before the 1980s was a major issue in education computer literacy.
* Not before the 1980s was computer literacy a major issue in education.

24. Categorization Task:

Raters must determine whether the four or more categories are logical and whether classification into these categories is consistent.

Scores are awarded on a 0-9 scale by giving a point credit for each logical classification, assessing a point penalty for each illogical or missing classification, dividing the total by 4, and rounding to the nearest whole number. Award a 0 score if the result is negative or if the categorization scheme is illogical on the whole.

28. Key:

A good comedy can be both entertaining and enlightening.
A good comedy can be both enlightening and entertaining.

30. Acceptable responses must make reasonable sense and be appropriately capitalized and punctuated.

Key:

Acceptable Responses:

Fish.
Fish have an extremely sensitive sense of smell.
Fish have sense.
Fish smell.
Fish smell extremely.
Have fish.
Have sense.
Smell.
Smell fish.

Unacceptable Responses:

Fish extremely.
Fish have smell.
Fish sense an extremely sensitive smell.
Fish smell sensitive.
Have fish sensitive smell.
Sense extremely.
Sense fish.
Sensitive fish have smell
Sensitive smell have fish.
Smell extremely.

Score on a 0-4 scale with 1 point for 1-2 acceptable responses, 2 points for 3-4 acceptable responses, 3 points for 5-6 acceptable responses, and 4 points for 7 or more acceptable responses. Deduct 1 point for 1-2 unacceptable responses, 2 points for 3-4 unacceptable responses, etc. If resulting score is less than 0, award a 0.

40. Key:

* As they age, people tend to get fewer colds. (comma optional)

* People as they age tend to get fewer colds.

* People tend as they age to get fewer colds.

* People tend to get fewer colds as they age.

44. Key: (3), (5), (1), (4), (2), where (3) indicates that the first sentence belongs in the third position in the paragraph.

Alternate Key: (2), (5), (3), (4), (1)

Score on a 0-5 scale by awarding 1 point for each correct placement of a sentence. <u>For imperfect responses only</u>, in addition to awarding points for absolute placement, grant 1/2 point for each correct sequence of two sentences. For example, the sequence (5), (1), (4), (2), (3), would receive 1.5 points (rounded to 2) for sequence: 1/2 point for 5 and 1, 1/2 for 1 and 4, and 1/2 for 4 and 2. Round all scores upward to the nearest whole number.

3. Substitution/Correction

Unless otherwise specified, items should be scored as "1" (correct) or "0" (incorrect).

5. Correct Solutions:

a) coming well after the discovery of "red giant" stars
b) coming well after that of "red giant" stars
c) occurring well after the discovery of "red giant" stars
d) occurring well after that of "red giant" stars
e) which came well after the discovery of "red giant" stars
f) which came well after that of "red giant" stars
g) which occurred well after the discovery of "red giant" stars
h) which occurred well after that of "red giant" stars
i) made well after that of "red giant" stars
j) made well after the discovery of"red giant" stars
k) which was made well after the discovery of "red giant" stars
l) which was made well after that of "red giant" stars
m) coming well after "red giant" stars were discovered

8. Key: Change "so" to "as" in line 2.

Many fans of Stephen King, the author of numerous popular horror novels,

assume that he is as mad as some of his characters.

25. Key:

LINE	
1.	science
3.	definition
4.	gardener
6.	plants
7.	disciplined
10.	helpful
11.	amateur
14.	February

Score on a 0-4 scale awarding 1/2 point for each corrected misspelling and subtracting 1/2 point for each originally correct spelling that is misspelled. Round up to the nearest integer. Award a 0 if the result is negative.

29. Key: Change "is" to "as"

The sixteenth-century art critic Vasari regarded the painting entitled the Mona Lisa as a wonderfully faithful reproduction of an actual person; to many nineteenth-century critics, it was a symbol to be decoded.

31. MANY CORRECT RESPONSES ARE POSSIBLE.

Responses should be scored as follows:

3: The response is a grammatical sentence that contains all of the original information.

Example: "The fires set to fumigate the houses of the victims of the Black Death destroyed many documents that could have identified those victims and their ancestors.

2: The response is a grammatical sentence that omits some of the original information.

Example: "Many of the victims and their ancestors could have been identified by documents that were destroyed by fires set to fumigate the houses."

OR

The response is a sentence with some grammatical or syntactical problem(s) that contains all of the original information.

Example: "Destroyed by fires set to fumigate the houses, many victims of the Black Death and their ancestors could have been identified by the documents."

1: The response is a sentence with some grammatical or syntactical problem(s) that omits some of the original information.

Example: "The victims and their ancestors could be identified by the documents, but fires set to fumigate the houses destroyed them.

0: The response is not a single sentence, or it is one marked by serious grammatical errors, incoherencies, and omissions of essential information.

Example: "To get fumigate from the Black Death many houses were burned and it destroyed many documents."

34. THERE ARE MULTIPLE CORRECT POSSIBILITIES.

Score on a 0-4 scale awarding 1/2 point for each acceptable substitution of a synonym for an underlined word or phrase. Round up to the nearest integer.

Examples: eatery ... restaurant
assuaged ... soothed
kindness ... hospitality

41. Key:

* The bobcat still roams the rocky outcrops of North America, though it is seldom seen or heard.
* The bobcat still, though it is seldom seen or heard, roams the rocky outcrops of North America.
* The bobcat, though it is seldom seem or heard, still roams the rocky outcrops of North America.

45. Key: Change "did" to "were"

The roads and means of transportation remain as they were thirty years ago; only the town hall with its television aerial is new.

46. Key: There are more than a dozen legitimate ways to do this.

Responses are scored as follows:

3: The response is a grammatical sentence that contains all of the original information.

Example: "Many fans of Stephen King, the author of numerous horror novels, assume that he is as crazy as some of his characters."

2: The response is a grammatical sentence that omits some of the original information.

Example "Many fans of his numerous horror novels assume that Stephen King is also crazy.

OR

The response is a sentence with some grammatical or syntactical problem(s) that contains all of the original information.

Example: "Stephen King is the author of numerous horror novels and is assumed by many of his fans that he is as crazy as some of his characters."

1: The response is a sentence with some grammatical or syntactical problem(s) that omits some of the original information.

Example: "Many fans assume that Stephen King, who is the author of numerous horror novels, and is also somewhat crazy."

0: The response is not a single sentence, or it is one marked by serious grammatical errors, incoherencies, and omissions of essential information.

Example: "Stephen King as author of horror novels, and crazy.

4. Completion

Unless otherwise specified, items should be scored as "1" (correct) or "0" (incorrect).

3. Score as 1 or 0.

Keys for lines 1 & 2:

what (direction)
the (direction)
the (direction) in which
the (direction) that
in what (direction)
which (direction)
in which (direction)
(direction) as
the (direction they are traveling) in
what (direction they are traveling) in
which (direction they are traveling) in

Key for line 3:

to (find)

6. Multiple Keys are possible:

"development," "evolution,"
"awakening, "growth,"

Or any noun that makes semantic sense.

17. Key: Many different completions are possible. A credited answer should convey the idea that property owners want to halt the dig, or begin construction of the new structure.

22. Key: Any participial form that makes semantic sense.
Responses are scored as follows:

2: a participial form that makes semantic sense
Example: "attempting"

1: a participial form that does not make proper semantic sense.
Example: "remembering"
OR
a non-participial form that makes semantic sense
Example: "eager"

0: a non-participial form that does not make proper semantic sense

27. Key: "in" after "delights"

The human mind delights in finding patterns -- so much so that we often mistake coincidence for profound meaning.

33. Key: "which"

35. Key: A correct answer can be a single sentence or multiple sentences as long as words are not modified, added, or deleted.

There are multiple correct possibilities in addition to the following:

This entire allegory, I said, you may now append, dear Glaucon, to the previous argument; the prison-house is the world of sight, the light of the fire is the sun, and you will not misapprehend me if you interpret the journey upwards to be the ascent of the soul into the intellectual world according to my poor belief, which at your desire, I have expressed -- rightly or wrongly God knows.

Score on a 0-6 scale awarding 1/2 point for each correctly inserted mark of punctuation and subtracting 1/4 point for each incorrectly inserted mark of punctuation. Round to the nearest integer and award a 0 if the result is negative.

39. Key:

Acceptable Responses:
* and
* as
* for
* since
* while

5. Construction

Unless otherwise specified, items should be scored as "1" (correct) or "0" (incorrect).

7. Write a Letter

Sum the total number of "yes" responses, divide by 3, and round to the nearest whole number. A yes/no decision is made for each feature noted below.

Information identifying the writer

1. Gives the correct name: Pat Carson
2. Gives the correct street address: 291 Westover Street
3. Gives the correct city: Tyland
4. Gives the correct state: CA or California
5. Gives the correct zip code: 99499

Information identifying the recipient

6. Gives the correct name of company: Tyland Training Center
7. Gives the correct address: Box 335
8. Gives the correct city: Tyland
9. Gives the correct state: CA or California
10. Gives the correct zip code: 99499

Date of letter

11. Gives the date the letter is being written
12. Places date in appropriate business letter position
13. Writes an appropriate greeting for a business letter
14. Punctuates the greeting according to business letter convention
15. Capitalizes the greeting correctly
16. Writes the greeting in an appropriate place

Business Letter Closing

17. Writes an appropriate closing for a business letter
18. Punctuates the closing according to business letter convention
19. Capitalizes the closing correctly
20. Writes the closing in an appropriate place

Reference to the advertisement

21. Names the newspaper: Golden News

22. Notes the date of the advertisement: Month, day, year

23. States the positions that will be opening in the categories

24. Describes the terms of the employment accurately

25. Notes the correct salary

The purpose for writing

26. States that he/she is applying for a position

27. Identifies the category (categories) he/she is applying for

The writer's qualifications

28 A - Gives some relevant facts or other background information about the writer's qualifications for a position

28 B - Gives substantial, relevant information about the writer's qualifications for a position

29 Gives additional information about the writer that may help persuade the recipient to accept the writer into the program

Use of language

30 Creates a respectful, business-like tone

31 A - Controls grammar and usage fairly well

31 B - Controls grammar and usage very well

32 A - Uses words accurately

32 B - Uses words effectively

33 Punctuates words correctly (e.g., uses apostrophes appropriately)

34 Capitalizes words correctly

Control of sentence structure

35 A - Generally forms simple sentences correctly

35 B - Generally forms simple and complex sentences correctly

35 C - Varies sentence structure effectively (to convey meaning)

36 A - Punctuates simple sentences correctly.

36 B - Puntuates simple sentences correctly and complex sentences fairly well.

36 C - Punctuates simple and complex sentences correctly.

14. "Gorilla Sightings" has an 8 point scoring guide:

<u>Content</u>: 4 points, one each for
* date of census
* location of gorillas
* number of gorillas in 1981
* number of gorillas in 1986

Writing: 4 points
4 = errorless
3 = 1 error in grammar, syntax, spelling, punctuation, word choice, or the coordination of sentences (if more than one sentence is given)
2 = 2 errors of sort described above
1 = 3 or more errors of the sort described above
0 = incoherent response, or not attempted

A sample "8" response is:

"A 1986 census recorded sightings of 280 gorillas in the Virunga Mountains of Rwanda, Zaire, and Uganda; this marks an increase from the 240 gorillas sighted in this same area in 1981."

20. Write an announcement

Score of 4 — A Successful Message

Gives all of the essential information
Presents the information clearly and concisely
Creates a positive tone
Is generally free of intrusive errors in spelling, grammar, and punctuation

Score of 3 — An Accurate Message

Gives all of the essential information
Presents the information in a way that makes no unnecessary demands on the reader, such as:

Embedding essential information in irrelevant information
Creating some confusion because of imprecise wording
Formatting the information in an inefficient or disorganized way
Having intrusive errors in spelling/grammar/punctuation

Score of 2 — A Fairly Accurate Message

Presents most of the essential information
States the information fairly clearly

Score of 1 — An Attempt to Convey a Message

Presents some of the essential information

Essential Information

Who:	Volunteers to help fix up Youth Center Mrs. Stone or Pat Carson may or may not be mentioned, as appropriate
What:	Fix roof and work on grounds Bring hammers, rakes, other appropriate tools Lunch provided
Where:	The Youth Center
When:	Saturday, March 21 8:30 - 12:00
Why:	To repair damage caused by storm