DOCUMENT RESUME

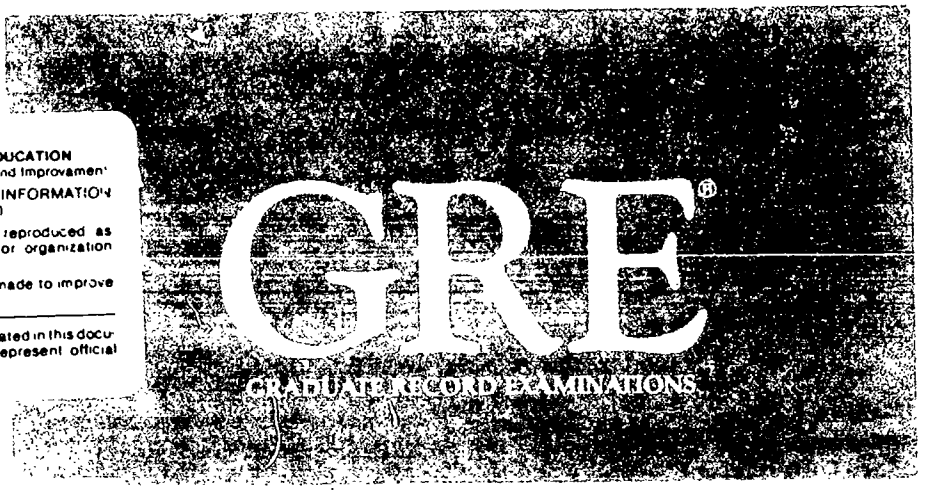
ED 388 721 TM 024 186

AUTHOR Chalifour, Clark; Powers, Donald E.
TITLE Content Characteristics of GRE Analytical Reasoning Items. GRE Board Professional Report No. 84-14P.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-88-7
PUB DATE May 88
NOTE 43p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Coding; *Difficulty Level; Higher Education; *Test Construction; Test Content; Test Items
IDENTIFIERS *Analytical Reasoning; *Graduate Record Examinations; Test Specifications

ABSTRACT

In actual test development practice, the number of test items that must be developed and pretested is typically greater, and sometimes much greater, than the number eventually judged suitable for use in operational test forms. This has proven to be especially true for analytical reasoning items, which currently form the bulk of the analytical ability measure of the Graduate Record Examination (GRE) General Test. This study involved coding the content characteristics of some 1,400 GRE analytical reasoning items and correlating them with indices of item difficulty, discrimination, and independence from the verbal and quantitative measures. Several item characteristics were predictive of the difficulty of analytical reasoning items. Generally, these same variables predicted item discrimination, but to a lesser degree. Independence from the GRE verbal and quantitative measures was largely unpredictable. The results suggest several content characteristics that could be considered in extending the current specifications for analytical reasoning items. The use of these item features may also contribute to greater efficiency in developing such items. Appendix A gives examples of analytical reasoning sets, and appendix B contains variable definitions and instructions to raters. (Contains 5 tables and 23 references.) (Author/SLD)



CONTENT CHARACTERISTICS OF GRE ANALYTICAL REASONING ITEMS

Clark Chalifour
and
Donald E. Powers

GRE Board Professional Report No. 84-14P
ETS Research Report 88-7

May 1988

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

ETS

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

Content Characteristics of GRE Analytical Reasoning Items

Clark Chalifour
Donald E. Powers

GRE Board Professional Report No. 84-14P

May 1988

Educational Testing Service, Princeton N.J. 08541

Acknowledgments

The authors greatly appreciate the help of Douglas Henson, Gilbert Harman, Elizabeth McGrail, and Eric Woisetschlaeger, who helped to develop the list of item characteristics; Peter Cooper, Elizabeth McGrail, and Gay MacQueen, who coded the items; Jennifer Kole and Inge Novatkoski, who performed the data analyses; and Walter Emmerich, Roy Freedle, and Eric Woisetschlaeger, who reviewed a draft of this report. Funding was provided by the Graduate Record Examinations (GRE) Board; members of the GRE Research Committee made helpful suggestions regarding several aspects of the study.

Abstract

In actual test development practice, the number of test items that must be developed and pretested is typically greater, and sometimes much greater, than the number that is eventually judged suitable for use in operational test forms. This has proven to be especially true for one item type--analytical reasoning--that currently forms the bulk of the analytical ability measure of the GRE General Test.

This study involved coding the content characteristics of some 1,400 GRE analytical reasoning items. These characteristics were correlated with indices of item difficulty, discrimination, and independence from the verbal and quantitative measures.

Several item characteristics were predictive of the difficulty of analytical reasoning items. Generally, these same variables also predicted item discrimination, but to a lesser degree. Independence from the GRE verbal and quantitative measures was largely unpredictable.

The results suggest several content characteristics that could be considered in extending the current specifications for analytical reasoning items. The use of these item features may also contribute to greater efficiency in developing such items. Finally, the influence of these various characteristics also provides a better understanding of the construct validity of the analytical reasoning item type.

> Ideally, test specifications should be so complete and so explicit that two test constructors operating from these specifications independently would produce comparable and interchangeable instruments, differing only in the sampling of questions included (Tinkelman, 1971, p. 47).

Unfortunately, this ideal is seldom if ever achieved. In actual test development practice, the number of test items that must be developed and tried out is typically (and usually significantly) greater than the number that is eventually judged suitable for use in operational test forms. Although this applies to most tests, it is especially true of the GRE analytical measure, whose item types continue to evolve. This state of affairs is consistent with research showing that even experienced test developers and subject matter experts may have difficulty in predicting the psychometric characteristics of test items (e.g., Bejar, 1983).

The goal of the study was to provide a better understanding of the characteristics of GRE analytical reasoning items and, as a result, to achieve greater precision and efficiency in the development of these test items. The specific objective was to identify content characteristics that are related to the difficulty and discriminating power of analytical reasoning items, as well as to their independence from GRE verbal and quantitative items.

The Analytical Reasoning Item Type

The focus was on the item type known as analytical reasoning, which currently forms the bulk of the GRE analytical measure. Logical reasoning, a less prominent item type used in the analytical measure, was not studied here. Analytical reasoning items occur in sets consisting of a stimulus passage followed by three to seven items based on the stimulus. Although stimuli are sometimes based on actual events, they most often describe fictitious situations created for the specific purpose of generating items with the desired characteristics. For the most part, the stimuli consist of related rules that govern such things as the allowable composition of a committee or the possible position of an event in a sequence. The items require the use of deductive reasoning to determine what is necessary or possible under the original set of rules or constraints, or under the original set plus one or more additional constraints introduced in the item stem. Rarely, an item requires the temporary suspension of a constraint. Examples of analytical reasoning items are shown in Appendix A.

All analytical reasoning items are pretested before being included in final test forms. Specifications for analytical reasoning items establish criteria for the level of difficulty and

discriminating power of the items. The relative independence of the verbal and quantitative measures is monitored through additional item analyses that use verbal and quantitative scores as criteria. (Analytical reasoning items are generally more highly correlated with the GRE quantitative measure than with the verbal measure.) Current test specifications also control the balance between items that ask what is necessarily true and items that ask what is possibly true. There are, however, no firmly established specifications for any other content characteristics.

Although the GRE verbal and quantitative measures are not curriculum-dependent tests, such information as the source and level of reading passages, and the educational level at which particular mathematical skills are taught, provide item writers with some clues to item difficulty. This kind of information is less relevant to analytical reasoning items, however. To gauge the difficulty of analytical reasoning items, test developers must rely almost exclusively on their own judgment and experience with the item types. Partly as a result of this, less experienced item writers tend to produce items that are either too easy or too difficult. Hence, a substantial surplus of items must be allowed in pretesting. If it were possible to use content characteristics to predict difficulty, discriminating power, and independence more precisely, content specifications could be refined or extended with a view toward maximizing the yield of items with desired measurement characteristics.

Relevant Research

As Glaser and his colleagues have pointed out (Glaser, 1986; Glaser & Lesgold, 1985), psychometric research has focused mainly on the end product of testing, i.e., test scores, almost to the exclusion of research on the development of tests. Consequently, relatively little research has been directed at helping test developers to better control the statistical characteristics of test items. Nonetheless, there has been some relevant research, scattered over a variety of test item types. For most of this research, the variable of primary interest has been item difficulty. Some studies have also examined effects of item characteristics on test reliability, test validity, and the discriminating power of items.

Verbal item types have been studied by several investigators. For example, Drum, Calfee, and Cook (1980) examined the influence of certain surface structure variables in reading comprehension items. Green (1984) studied the effects of sentence length, syntactical complexity, and the use of uncommon words. Kirsch and Guthrie (1980) reported on some factors that affected the difficulty of functional literacy items. Bejar (1983) found that subject matter experts were unable to discover factors that may contribute to the difficulty of a test of conventional and standard written English. Freedle and

Fellbaum (1987) investigated the role of several language features in the difficulty of listening comprehension items, finding that lexical repetition is an important factor in item difficulty. Mitchell (1983) found that the difficulty of word knowledge items was predicted by the amount of information presented and the usage frequency of words in the items.

Nonverbal item types have also been studied. For instance, Campbell (1961) looked at the nature of the classifying concepts (e.g., shape and size) in nonverbal classification tests. Bejar (1986) found that angular disparity was a potent determinant of item difficulty in a three-dimensional spatial task. Mulholland, Pellegrino, and Glaser (1980) identified two process factors--encoding complexity and transformation complexity--that were quite predictive of the difficulty of geometric analogies.

A variety of other research has investigated the relationship of item difficulty to numerous structural characteristics of items. A number of these studies have concerned the effects of violating accepted principles of item writing (e.g., Board & Whitney, 1972; McMorris, Brown, Snyder, & Pruzek, 1972) or of using alternative item formats (e.g., Dudycha & Carpenter, 1973; Green, 1984; Hughes & Trimble, 1965).

The research cited above provides some general clues regarding the characteristics that affect the difficulty of test items, but it does not relate specifically to measuring analytical abilities. Research that relates somewhat more directly to the analytical reasoning item type is the work of Dickstein (1976), Guyote and Sternberg (1981), and Nester and Colberg (1984). Each of these studies focused on factors involved in testing syllogistic reasoning, but this is a much more specific deductive reasoning skill than is involved in solving GRE analytical reasoning items. Nonetheless, these studies do suggest the possible importance of such variables as (a) verbal form (factual syllogisms are easier than either counterfactual or anomalous ones, Guyote and Sternberg, 1981), and (b) linguistic medium (symbolic vs. verbal, Nester & Colberg, 1984). Formal logical properties like negation mode, i.e., the use of negative equivalents of logical propositions, have also been shown to affect the difficulty of syllogistic items (Nester & Colberg, 1984).

Method

The first phase of the study consisted of specifying content characteristics that were likely to be of interest. An initial list was prepared by the person who served for 10 years as primary test developer for the analytical measure. This list was reviewed by other experienced test development staff and college teachers of logic who had previously served as reviewers of analytical reasoning items. No content characteristics were added as the result of these reviews, but

several were eliminated because they could not be identified reliably even by experienced reviewers.

The final list of characteristics, attached as Appendix B, included 17 characteristics that pertain to the stimulus or the set as a whole and 10 characteristics that describe each of the individual items in a set. The identification of 5 of the 17 stimulus/set characteristics and 2 of the 9 individual item characteristics required expert judgment or actual solution of the items; the remaining characteristics (12 stimulus and 7 item features) could be identified by trained clerical staff. Responses consisted of numerical codes indicating judgmental ratings (e.g., of the usefulness of simple diagrams as an aid to answering questions); actual quantities (e.g., the number of words in a stimulus); or the presence or absence of specific features (e.g., the presence of the word "must" in options). One of the judgments was the expert's estimate of the percentage of test takers who would correctly answer each item. These percentages were converted to delta values, the standard index used in ETS item analyses. All items were coded for 7 characteristics by an expert judge and for 19 other characteristics by a clerical coder. To ensure that the judgments of the expert were not idiosyncratic, her responses to a representative sample of four stimuli and 20 items were compared with the responses of three other experts with similar qualifications. For the clerical coding, a 100% quality control check was carried out for a sample of 24 sets of items. The characteristics of 227 analytical reasoning sets (1,474 items) in all were coded for this study. These included sets that were used in final test forms, as well as sets that did not meet operational specifications.

The primary method of analysis was multiple regression analysis using each of several dependent variables. The dependent variables were

(1) item difficulty, as reflected by the delta index, a normalized transformation of the percentage answering each item correctly

(2) item discrimination, as reflected by biserial correlations of each item with a total analytical score (transformed to Fisher's z) and

(3) independence, as reflected by the differences between the biserials based on the analytical score and those based on either the verbal score or the quantitative score

The variables describing stimulus and item characteristics served as the independent variables in the analyses. (With dummy variable coding of several nominal variables, the actual number of independent variables was 39.)

A stepwise procedure was employed in which variables were added to the prediction system until no variable added significantly to prediction. (It should be noted that the intercorrelations among variables were generally quite low so that collinearity was not a major problem.) Two different stopping rules were employed. One terminated the selection of variables when no statistically significant increase (at the .05 level) resulted. The other rule terminated the addition of variables when the R^2 value did not increase by at least .01. Thus, a set of the most predictive variables was identified for each dependent variable, as was a smaller subset based on a second, more stringent variable selection rule. These analyses were run on approximately 1,300 items (90% of the total items that were coded). The remaining items, a 10% sample of all coded items, were withheld from these analyses so the results could be cross-validated.

Results

Reliability of Classifications

A 100% quality control of the clerical coding of 24 items revealed five characteristics that were problem areas. The ratings for each of these characteristics were checked and corrected as necessary for all items by the senior author.

Table 1 displays various indices of agreement among the four expert judges who rated, for a subset of four stimuli and 26 items, the seven more subjective characteristics that required expert judgment. The focus here was on assessing, before the bulk of items was coded, the degree to which the primary expert rater might provide idiosyncratic ratings. The indices reflect the degree to which the four judges agreed exactly (summed over all comparisons between each pair of judges) and the extent to which their agreement was at a maximum. Maximum disagreement was defined as an instance in which one judge gave the highest possible rating and another the lowest rating on the scale. Correlations are also given between the ratings made by the primary expert judge (judge 1) and each of the other three judges, and between the primary judge and the average of ratings given by the other three judges.

As is apparent, except for agreement about "degree of realism," the various indices suggest moderately good agreement about each of the other six characteristics. Ratings involving the next least agreed upon characteristic--the amount of information that must be used to solve the problem--were made relatively reliably, with 59% of all judgments corresponding exactly.

Description of Stimulus/Item Characteristics

Table 1 shows the characteristics of the almost 1,500 analytical items that were coded for this study. For most of the characteristics, there was at least moderate variability among items or stimuli For example, items are almost exactly evenly split with regard to whether they ask "What must be true?" versus "What can be true?" This was expected, because this feature is currently the primary test specification for analytical reasoning items. Other characteristics exhibited considerably less variation over items. At one extreme, for instance, nearly all the items refer to objects by names, numbers, or symbols instead of by relational terms.

The items coded for this study also showed variation in terms of difficulty and discrimination, as shown in Table 2. Item difficulty was, on average, slightly greater and discrimination somewhat less for the items studied here than for items that are typically included in final operational test forms. These discrepancies result from the inclusion in this study of items that did not meet the specifications for final test forms.

Relationships among Stimulus/Item Characteristics

The degree to which the various characteristics of stimuli and items are related to one another is important in two respects. First, the stability of results from our primary method of analyzing data--stepwise multiple regression analysis--depends on the relative independence of our explanatory variables from one another. Second, the extent to which these variables are interrelated may also suggest the degree to which these variables may be amenable to independent manipulation when constructing analytical reasoning items.

Generally, the correlations among variables were relatively low (less than .20 in absolute value). An inspection of the intercorrelation matrix revealed several clusters of variables that seem to relate both logically and empirically. These patterns of correlations seem to suggest several dimensions. For example, an "information load" dimension was suggested by correlations of .64, .56, and .37 among the "number of sentences in the stimulus," the "number of words in the stimulus," and the "number of rules, conditions, or restrictions included." High correlations were also noted between the use of various standard item stems, e.g., "which of the following statements must be true?", and the nature of options, e.g., as lists of names, or as statements. The highest correlations among the several subparts of these variables were -.59, .70, and .84. No other prominent clusters of correlations were readily apparent.

Prediction of Statistical Characteristics

The major results of the regression analyses are shown in Table 3. This table displays the regression equations for predicting item difficulty, both actual and estimated, and item discrimination. The prediction of item independence is also shown under item discrimination. Two indices of independence were used--one based on the difference between biserials when an analytical criterion versus a verbal criterion was used and again when an analytical versus a quantitative criterion was used. These indices of independence reflect the degree to which an item correlated relatively better with the analytical measure than with the verbal (or quantitative) measure.

Two equations are shown for each of these dependent variables. Equation 1 is based on the selection of variables according to whether or not they added significantly to prediction ($p < .05$) in combination with the variables selected previously. Equation 2 is based on a smaller set of predictors that added at least .01 to the R^2 value resulting from previously entered variables. The weights shown are raw regression weights. The values of weights that did not add significantly to prediction have been omitted.

Item difficulty. To predict item difficulty, a model based on 18 of the 39 variables was specified when the first variable selection rule was employed. The resulting multiple R was .65. In contrast, the estimates of item difficulty provided by the expert judge (an experienced item writer) correlated .72 with actual item difficulty when expressed on the same normalized scale. A smaller subset of seven of these variables, selected by the second rule, did almost as well, with a multiple R of .62. The smaller set of item or stimulus characteristics that contributed to this correlation were

1. usefulness of drawing diagrams in reaching a solution (the greater the usefulness of, or need for, diagrams, the more difficult the item)

2. number of words in the stimulus (the more words, the more difficult)

3. number of unvarying assignments to position (the more, the easier)

4. number of rules or conditions (the more rules, the more difficult)

5. amount of information from the rules or conditions that is needed to solve the problem (the more, the more difficult)

6. use of the item stem "which is a possible sequence..." (easier than other stems)

7. options that are lists of names instead of numbers, or statements (easier than other options)

The comparison of the prediction of estimated difficulty with the prediction of actual difficulty shows some interesting differences. Generally, the variables that contributed to the prediction of actual item difficulty also predicted estimates of item difficulty. When the smaller subsets of predictors are compared (equation 2), it can be seen that five of the same variables were predictive of both actual and estimated difficulties. Two additional variables that predicted actual difficulty did not significantly predict estimated difficulty. These were (a) the number of unvarying assignments and (b) the number of rules or conditions. One possible implication of this difference is that item writers may be less inclined to notice these variables or may consider them as less important than others. Two other variables--(a) the degree to which diagrams prove useful and (b) the extent to which options are lists--were weighted more heavily in the prediction of actual than estimated difficulty.

Discrimination. To predict discriminating power, a model based on 14 of the 39 variables was specified. The resulting multiple R was .46, substantially less than that obtained for the prediction of difficulty. A six-variable subset did almost as well, with a multiple R of .41. Generally, the variables selected in the prediction of item difficulty were also selected here. However, because of the strong correlation between item difficulty and biserial correlations (-.66), the characteristics that were associated with higher difficulties were usually associated with lower biserial correlations (discrimination indices). There were some exceptions. The number of subclassifications or subgroupings mentioned in the problem was related to item discrimination, but not to item difficulty (the more subclassifications, the lower the discrimination). Two characteristics--the number of unvarying assignments and the number of rules or conditions--were related to item difficulty but not to discrimination. Six of the seven characteristics that predicted item difficulty in the large sample were selected again in the stepwise analyses in the smaller sample, as were four of the six characteristics that predicted discriminability. When the regression weights computed for the larger (90%) sample were applied in the smaller 10% sample, the resulting multiple R values decreased from .65 and .62 to .46 and .45 for item difficulty and from .46 and .41 to .32 and .33 for discrimination, using the full models and the subsets, respectively.

Independence. The prediction of independence, i.e., the difference between the biserials based on an analytical score versus those based on either a verbal score or a quantitative score, was slight. This result was not unexpected, given the very high correlations among biserials when analytical, verbal, and quantitative criteria were used. The correlation between analytical and verbal biserials was .92; that between analytical and quantitative biserials was .91. The two measures of independence therefore reflected little

that was unique to analytical biserials. A set of 8 variables yielded a multiple R of .27 for the prediction of the difference between the analytical and verbal biserials, while a set of 15 variables yielded a multiple R of .42 for predicting the difference between analytical and quantitative biserials. Smaller sets of predictors selected on the basis of a .01 contribution to R^2 had multiple Rs of only .16 and .35. The multiple Rs in the cross-validation analyses ranged from .06 to .29 for the four equations.

Relative importance of predictors. Table 4 shows only the raw regression weights for each variable and indicates the actual weights that would be applied to the value of each variable to predict each of the dependent variables--difficulty, discrimination, and independence. Standardized weights, on the other hand, indicate the relative importance of each variable in a combination of predictors. An examination of these weights (Table 5) shows that two variables--(a) the extent to which diagrams are judged to be useful and (b) the number of words in the stimulus--contribute most to the prediction of both item difficulty and item discrimination. Each of three other variables--(a) the use of options as lists of names, etc., (b) the use of "which is a possible sequence...," and (c) the amount of information to be used--contributed to a lesser degree. Two other variables--the number of unvarying assignments and the number of rules or conditions--were somewhat less important though still significant. The five most influential of these variables contributed to about the same degree to the prediction of item discrimination.

Table 5 suggests that decreasing the number of unvarying assignments might be one way to make items more difficult without concomitantly decreasing discrimination; increasing the number of rules or conditions might also increase difficulty without impairing discrimination. However, although not selected with the small subset of variables shown in Table 4, the weights computed for these two variables were not appreciably less than the weights of the variables listed in Table 4. Decreasing the number of subgroups or subclassifications, on the other hand, might be one way to increase item discrimination without altering difficulty level, because this variable was predictive of item discrimination but received a small and nonsignificant weight for predicting item difficulty. As is apparent from Table 1, however, relatively few subgroups are used in analytical reasoning items--on average, only 0.3.

Controlling for difficulty. Because of the strong relationships between indices of difficulty and discrimination, a further regression analysis was run with discrimination as the dependent variable. In this analysis, item difficulty was first partialled out. This analysis revealed that only one characteristic--whether objects in the stimulus were identified by names, numbers, or other symbols or whether they were identified in relational terms--was predictive of item discrimination when difficulty was held constant. The multiple R

was .79, and no other variable contributed more than .01 to this correlation when added to this variable.

When the same analysis was applied to the prediction of independence, the same variable was also the most significant predictor of relative independence from the verbal measure and from the quantitative measure. The multiple Rs for predicting independence of the verbal and quantitative measures were .49 and .52 from this variable. One additional variable added significantly to the R^2 for the prediction of independence from the verbal measure. Three other additional variables contributed significantly to the prediction of independence from the quantitative measure. Each of these variables, however, was barely significant. In relation to the single most predictive variable, none made a practically significant contribution to prediction.

Summary and Implications

A number of characteristics of analytical reasoning items, the most prominent item type in the GRE analytical measure, were identified. Heretofore, most of these characteristics have not been considered, at least formally, in the development of analytical reasoning items. With one exception, these characteristics can be reliably coded, many of them by trained clerical staff.

Several item characteristics were shown to predict the difficulty of analytical reasoning items. The strongest predictors were (a) the degree to which drawing a diagram proves useful, (b) the number of words in the stimulus, (c) the use of the item stem "which is a possible sequence...," (d) whether or not options are lists of names..., and (e) the amount of information to be used in solving the problem. These variables are capable of accounting for about 30% of the variation in item difficulty and therefore might be strong candidates for extending the specifications for the analytical reasoning item type. Interestingly, the current major specification for analytical reasoning items--whether a statement is necessarily true versus possibly true--did not contribute significantly to the prediction of item difficulty when considered in combination with other variables. This variable did, however, add significantly to the prediction of discriminability when a larger set of predictors was used.

Generally, the same variables that predicted item difficulty were also predictive of item discrimination but to a lesser degree, accounting for about 10-15% of the variation. Because of the strong negative relationship of item difficulty to item discrimination, each variable that was associated with greater difficulty also forecast

less discrimination.[1] Despite the reason(s) for this negative relationship, it clearly suggests the tradeoff that must be considered in developing these items. The detailed results of this study sometimes pointed to item characteristics that, when varied, might have a greater impact on difficulty than discrimination, or vice versa. However, these particular variables were generally relatively weak predictors of either difficulty or discrimination.

This study also examined the degree to which the various item characteristics predicted independence, i.e., the difference between the relationship of analytical items to an analytical criterion and their relationship to a verbal (or quantitative) criterion. The results revealed little to suggest the possibility of making these predictions. Forecasts of this nature were precluded largely because of the very strong correlation between discrimination indices based on verbal, quantitative, and analytical criteria. If an item related strongly to one criterion, it tended to relate strongly to others.

With respect to estimating item difficulty, it appears, on the basis of a limited sample, that experienced item writers are capable of estimating the difficulty of analytical reasoning items. The relationship between actual difficulty and one expert's estimates of difficulty was relatively strong, especially in comparison with results obtained for other item types studied in previous research. More important, this study suggested that, because of their differential relationship to actual and estimated difficulty, some important item characteristics may not be noticed by test developers, and other item features may be accorded more weight than is warranted. This speculation could be evaluated by test development staff.

The results of this study may be important also with respect to implications regarding the construct validity of the analytical reasoning item type. The two most influential predictors of item difficulty may have somewhat different implications. On the one hand, the degree to which a diagram proves useful in solving analytical reasoning items was important. This would seem to relate to the construct described in the current GRE Information Bulletin

---

[1]This situation is encountered relatively frequently in test development and, in fact, is one of the major constraints on objective testing. There may be many factors involved in this negative relationship. One possibility is that items may be difficult because they reflect not only knowledge or abilities that are intended to be measured, but also abilities that may be less directly relevant. To the extent that difficult items are more reflective of these unintended sources of variance, they can be expected to correlate less strongly with a criterion that reflects, primarily, the intended abilities. A more thorough consideration of this situation, although warranted, is beyond the scope of the study reported here.

(Educational Testing Service, 1987). The analytical reasoning item is characterized as testing

> ...the ability to understand a given structure of arbitrary relationships among fictitious persons, places, things, or events... (p. 37).

Furthermore, test takers are advised that to understand these structures it may be useful to draw rough diagrams. The study results would seem, at least on one level, to support these statements. However, further thought regarding this finding, in relation to the psychological meaning of this predictor, may be needed. As Emmerich (personal communication) has suggested, one plausible alternative interpretation of the importance of the "usefulness of drawing diagrams" is that raters may have first judged an item's difficulty using various other cues and then judged that, for the more difficult items, drawing a diagram would be especially useful. Within the limits of our data, there is no way to discount this rival hypothesis.

On the other hand, the verbal load, as indexed by the number of words in the stimulus, was nearly as important as the usefulness of a diagram. This finding, although less comforting, may suggest a useful course of action, such as controlling the variability among analytical reasoning stimulus sets with respect to length, and perhaps keeping the number of words to a minimum. The implications regarding the other strong predictors of item difficulty do not seem as readily apparent, either positively or negatively.

Finally, we note the limitations of the study reported here. First, the study focused mainly on the surface characteristics of GRE analytical reasoning items, not on the psychological processes that underlie the solution of these items. This strategy was thought to be appropriate given the current stage of evolution of analytical reasoning items. Moreover, these surface features may relate to, or interact with, the more meaningful underlying processes. It is highly likely that the development of analytical reasoning items would benefit further from research on these processes. Cognitive research on verbal analogies has resulted in a rich and established literature and provides a good example. This literature is currently being applied to understanding, and eventually to developing, the verbal analogy items used in the GRE General Test (Bejar, Embretson, Peirce, & Wild, 1985).

Second, the model assumed here was a strictly linear one. Given the relatively large number of variables (even in relationship to the large number of items), this simplifying assumption was deemed to be desirable. It is quite possible, however, that the various characteristics of items and stimuli may interact in complex ways, and the presence of two or more characteristics may contribute more than either one alone to the prediction of item difficulty and discrimination. Future studies exploring these interactions would

presumably be facilitated by the results of the current study, especially with respect to reducing the number of variables of interest. Finally, although the results suggest some potentially useful ways to improve the efficiency with which analytical reasoning items are developed, true experimental studies are needed to determine the actual effects of the item writing strategies implied here. Such studies are needed to ensure that the variables investigated here are in fact amenable to manipulation without having any unintended negative effects.

In summary, we hope the results of this study may contribute in a modest way to some of the benefits discussed by Scheuneman and Steinhaus (1987): fewer items lost in pretesting, more precisely delineated content specifications, more rational defense of individual items when challenges occur, and improved construct validity. We suggest that the primarily empirical approach employed here might benefit in any future studies of this kind from a more comprehensive theoretical framework, such as the one suggested by Scheuneman and Steinhaus (1987), in which not only the demands of test items are considered but also the characteristics of examinees and the interaction between the characteristics of items and those of examinees.

References

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. Applied Psychological Measurement, 7, 303-310.

Bejar, I. I. (1986). A psychometric analysis of a three-dimensional spatial task (ETS Research Report RR-86-19-ONR). Princeton, NJ: Educational Testing Service.

Bejar, I., Embretson, S., Peirce, L., & Wild, C. (1985). Application of cognitive research for GRE test development. Proposal to the GRE Research Committee, January 1985.

Board, C., & Whitney, D. R. (1972). The effect of selected poor item writing practices on test difficulty, reliability, and validity. Journal of Educational Measurement, 9, 225-233.

Campbell, A. C. (1961). Some determinants of the difficulty of nonverbal classification items. Educational and Psychological Measurement, 21, 899-913.

Dickstein, L. S. (1976). Differential difficulty of categorical syllogisms. Bulletin of the Psychonomic Society, 8, 330-332.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1980). The effect of surface structure variables on performance in reading comprehension tests. Reading Research Quarterly, 16, 486-513.

Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. Journal of Applied Psychology, 58, 116-121.

Educational Testing Service (1987). GRE Information Bulletin. Princeton, NJ: Author.

Freedle, R. O., & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In R. O. Freedle and R. P. Duran (Eds.), Cognitive and linguistic analyses of test performance. Norwood, NJ: Ablex Publishing Corporation.

Glaser, R. (1986). The integration of instruction and testing. In The redesign of testing for the 21st century (proceedings of the 1985 ETS Invitational Conference). Princeton, NJ: Educational Testing Service.

Glaser, R., & Lesgold, A. (April 1985). Cognitive task analysis and the measurement of proficiency. Paper presented at the Buros-Nebraska Symposium on Measurement and Testing, Lincoln, NE.

Green, K. (1984). Effects of item characteristics on multiple choice item difficulty. Educational and Psychological Measurement, 44, 551-562.

Guyote, M. J., & Sternberg, R. J. (1981). A transitive chain theory of syllogistic reasoning. Cognitive Psychology, 13, 461-525.

Huck, S. W., & Bowers, N. D. (1972). Item difficulty level and sequence effects in multiple-choice achievement tests. Journal of Educational Measurement, 9, 102-111.

Hughes, H. H., & Trimble, W. E. (1965). The use of complex alternatives in multiple choice items. Educational and Psychological Measurement, 25, 117-126.

Kirsch, I. S., & Guthrie, J. T. (1980). Construct validity of functional reading tests. Journal of Educational Measurement, 17, 81-93.

McMorris, R. F., Brown, J. A., Snyder, G. W., & Pruzek, R. M. (1972). Effects of violating item construction principles. Journal of Educational Measurement, 9, 287-295.

Mitchell, K. T. (1983). Cognitive processing determinants of item difficulty on verbal subtests of the Armed Services Vocational Aptitude Battery (Technical Report 598). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Mulholland, T., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. Cognitive Psychology, 12, 252-284.

Nester, M. A., & Colberg, M. (1984). The effects of negation mode, syllogistic invalidity, and linguistic medium on the psychometric properties of deductive reasoning tests. Applied Psychological Measurement, 37, 91-100.

Scheuneman, J. D., & Steinhaus, K. S. (1987). A theoretical framework for the study of item difficulty and discrimination. Princeton, NJ: Educational Testing Service.

Tinkelman, S. N. (1971). Planning the objective test. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education.

Table 1

Reliability of Expert Judgments

Variable	Percentage of: Perfect agreement over all judges	Percentage of: Maximum disagreement over all judges	Correlation between Judges: 1 & 2	1 & 3	1 & 4	1 & average of 2, 3, 4
Usefulness of drawings, charts, or diagrams	67	9	.30	.85	.97	.74
Number of possible configurations of entities	100	0	1.00	1.00	1.00	1.00
Degree of realism of problem	17	0	.87	-.64	.30	.01
Relation of stimulus to an academic discipline	92	n.a.	1.00	1.00	.58	1.00
Kind of task	75	n.a.	.43	1.00	.82	.89
Item classification	94	n.a.	.92	.92	.61	.98
Amount of information that must be used	59	0	.62	.59	.40	.74

n.a. = not applicable because variable was categorical

Table 2

Description of Analytical Reasoning Items (N = 1,474 items)

Characteristic	Scale	Mean or Percentage of Occurrence (S.D)
Stimulus characteristics		
Usefulness of drawings, charts, or diagrams	1 = not useful to 5 = useful	3.39 (1.58)
Number of possible configurations of entities	1 - 10, 11 if more than 10	9.9 (2.3)
Degree of realism of problem	1 = unrealistic to 5 = realistic	2.83 (1.03)
Relation of stimulus to an academic discipline	Not related Related	90.2% 9.8%
Kind of task	Ordering Determination of set membership Combination Other	35.3% 30.5% 12.4% 21.8%
Number of sentences in stimulus		7.1 (2.1)
Number of words in stimulus		112 (35)
Number of persons, objects, etc., to be ordered		6.7 (1.7)
Number of subclassifications or subgroupings		0.3 (0.6)
Number of unvarying assignments		0.3 (0.6)
Number of positions in any orderings or groups		6.4 (2.2)
Number of rules, conditions, or restrictions		4.4 (1.4)
Degree of use of terminology of math or formal logic	Number of occurrences of various phrases such as "if and only if"	0.4 (0.7)
Number of simultaneous configurations, orderings, or groupings		1.2 (0.6)

Table 2 (Continued)

Characteristic	Scale	Mean or Percentage of Occurrence (S.D)
Method of labeling objects	Names	27.0%
	Letters	66.0%
	Numbers	3.9%
	Other attributes	3.1%
Composition of the pool of objects	Living things	48.1%
	Inanimate objects	41.6%
	Other	10.3%
References to objects	Identified by name, numbers, or symbols	99.6%
	Identified in relational terms	0.4%
Item Characteristics		
Item classification	Asks what must be true	49.2%
	Asks what can be true or what is not possible	50.8%
Amount of information that must be used	1 = none to	4.3
	5 = all	(1.0)
Type of item stem	"Which statement must be true?"	18.1%
	"Which of the following could be true?"	3.3
	"Which of the following is a complete and accurate list?"	2.0%
	"Which is the greatest [least] number of...?"	6.2%
	"Which of the following is a possible sequence, ordering, etc.?"	9.3%
	Other	61.1%
Whether stem adds new conditions	No new conditions	44.6%
	New conditions added	55.4%
Whether or not stem suspends any original conditions	No original conditions suspended	99.7%
	Some original conditions suspended	0.3%

Table 2 (Continued)

Characteristic	Scale	Mean or Percentage of Occurrence (S.D)
Whether or not stem asks for negative response	No negative response	84.9%
	Negative response	15.1%
Nature of options	Lists	61.9%
	Numbers	8.0%
	Positive statements	23.9%
	Negative statements	0.1%
	Positive and negative statements	3.6%
	Other	2.5%

Table 3

Psychometric Characteristics of Analytical Reasoning Items (N = 1,474)

Characteristic	Mean	S.D.
Difficulty		
Actual (delta)	12.9	2.8
Estimated (% correct)	55.6	16.7
Discrimination (biserial r)		
Analytical criterion	.37	.17
Verbal criterion	.25	.14
Quantitative criterion	.30	.15

Table 4

Predictions of Statistical Properties from Item/Stimulus Characteristics (N = 1,326)

Item or Stimulus Characteristic		Item Difficulty: Estimated Eq.1	Item Difficulty: Estimated Eq.2	Item Difficulty: Actual Eq.1	Item Difficulty: Actual Eq.2	Item Discrimination: Analytical* biserial Eq.1	Item Discrimination: Analytical* biserial Eq.2	Item Discrimination: Analytical* vs. Verbal Eq.1	Item Discrimination: Analytical* vs. Verbal Eq.2	Item Discrimination: Analytical* vs. Quantitative Eq.1	Item Discrimination: Analytical* vs. Quantitative Eq.2
Intercept	b	6.9	7.7	5.7	6.5	5.1	6.1	0.9	0.7	1.5	1.5
	(SE)	(.2?)	(.25)	(.53)	(.42)	(.31)	(.25)	(.12)	(.07)	(.16)	(.09)
Stimulus Characteristics											
1. Usefulness of diagrams (1 = not useful, 5 = useful)	b	.27	.28	.44	.54	-.21	-.21	---	---	-.08	-.08
	(SE)	(.03)	(.03)	(.05)	(.04)	(.03)	(.03)	---	---	(.01)	(.01)
3. Realism of problem (1 = unrealistic, 5 = realistic)	b	---	---	-.25	---	.08	---	---	---	---	---
	(SE)	---	---	(.07)	---	(.04)	---	---	---	---	---
4. Relationship to a particular discipline (1 = related, 0 = unrelated)	b	---	---	.60	---	---	---	---	---	---	---
	(SE)	---	---	(.25)	---	---	---	---	---	---	---
5A. Task requires ordering	b	-.49	---	---	---	---	---	---	---	.20	---
	(SE)	(.10)	---	---	---	---	---	---	---	(.05)	---
5B. Task requires determination of set membership	b	-.25	---	---	---	---	---	---	---	.25	---
	(SE)	(.11)	---	---	---	---	---	---	---	(.05)	---
5C. Task requires combination of ordering and determination of membership	b	---	---	.44	---	---	---	.15	---	.31	---
	(SE)	---	---	(.20)	---	---	---	(.06)	---	(.06)	---
6. Number of sentences in stimulus	b	---	---	---	---	---	---	---	---	---	---
	(SE)	---	---	---	---	---	---	---	---	---	---
7. Number of words in stimulus	b	.01	.01	.02	.02	-.01	-.01	---	---	---	---
	(SE)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)	---	---	---	---
8. Number of objects to be arranged, ordered, etc.	b	---	---	---	---	---	---	.06	.06	.04	---
	(SE)	---	---	---	---	---	---	(.01)	(.01)	(.01)	---
9. Number of subgroups	b	---	---	---	---	-.27	-.35	---	---	---	---
	(SE)	---	---	---	---	(.07)	(.08)	---	---	---	---

Table 4 (Continued)

Item or Stimulus Characteristic		Item Difficulty: Estimated Eq.1	Item Difficulty: Estimated Eq.2	Item Difficulty: Actual Eq.1	Item Difficulty: Actual Eq.2	Item Discrimination: Analytical* biserial Eq.1	Item Discrimination: Analytical* biserial Eq.2	Item Discrimination: Analytical* vs. Verbal Eq.1	Item Discrimination: Analytical* vs. Verbal Eq.2	Item Discrimination: Analytical* vs. Quantitative Eq.1	Item Discrimination: Analytical* vs. Quantitative Eq.2
10. Number of unvarying assignments	b	-.30	---	-.64	-.65	.22	---	.10	---	---	---
	(SE)	(.07)	---	(.11)	(.11)	(.06)	---	(.03)	---	---	---
11. Number of positions to be developed	b	.07	---	.08	---	---	---	---	---	-.03	---
	(SE)	(.02)	---	(.03)	---	---	---	---	---	(.01)	---
12. Number of rules or conditions	b	.12	---	.27	.27	-.12	---	---	---	---	---
	(SE)	(.03)	---	(.05)	(.05)	(.03)	---	---	---	---	---
13. Use of math or logic terminology	b	---	---	---	---	---	---	---	---	.06	---
	(SE)	---	---	---	---	---	---	---	---	(.03)	---
15A. Labels are persons or things	b	---	---	-.74	---	---	---	---	---	-.31	---
	(SE)	---	---	(.28)	---	---	---	---	---	(.11)	---
15B. Labels are letters of alphabet	b	---	---	-.94	---	.24	---	.12	---	-.49	-.16
	(SE)	---	---	(.26)	---	(.08)	---	(.04)	---	(.11)	(.04)
15C. Labels are numbers	b	---	---	---	---	---	---	---	---	-.50	---
	(SE)	---	---	---	---	---	---	---	---	(.14)	---
16A. Pool composed of living things	b	.23	---	1.58	---	---	---	---	---	---	---
	(SE)	(.08)	---	(.25)	---	---	---	---	---	---	---
16B. Pool composed of inanimate objects	b	---	---	1.27	---	.24	---	---	---	---	---
	(SE)	---	---	(.24)	---	(.08)	---	---	---	---	---
Item Characteristics											
18. Item classification (necessarily true vs. can be true)	b	---	---	---	---	.19	---	---	---	---	---
	(SE)	---	---	---	---	(.09)	---	---	---	---	---
19. Amount of information to be used (1 = none, 5 = all)	b	.67	.61	.60	.52	-.24	-.21	-.06	---	-.12	-.12
	(SE)	(.04)	(.04)	(.06)	(.07)	(.04)	(.04)	(.02)	---	(.02)	(.02)

Table 4 (Continued)

Item or Stimulus Characteristic		Item Difficulty: Estimated Eq.1	Estimated Eq.2	Actual Eq.1	Actual Eq.2	Item Discrimination: Analytical* biserial Eq.1	Analytical* biserial Eq.2	Analytical* vs. Verbal Eq.1	Analytical* vs. Verbal Eq.2	Analytical* vs. Quantitative Eq.1	Analytical* vs. Quantitative Eq.2
21B. Use of stem "which could be true"	b	---	---	---	---	---	---	.29	---	.22	---
	(SE)	---	---	---	---	---	---	(.10)	---	(.10)	---
21C. Use of stem "which is complete list..."	b	.78	---	---	---	---	---	---	---	.33	---
	(SE)	(.29)	---	---	---	---	---	---	---	(.13)	---
21E. Use of stem "which is a possible sequence, etc."	b	-1.55	-1.85	-1.86	-2.16	.47	.64	---	---	.38	.42
	(SE)	(.15)	(.15)	(.23)	(.22)	(.14)	(.15)	---	---	(.06)	(.06)
22. Stem adds new conditions	b	.29	---	.31	---	---	---	---	---	---	---
	(SE)	(.09)	---	(.14)	---	---	---	---	---	---	---
24. Stem asks for negative response	b	(.42)	---	.75	---	-.34	---	---	---	---	---
	(SE)	(.12)	---	(.18)	---	(.12)	---	---	---	---	---
25A. Options are lists of names, etc.	b	-.49	-.50	-1.14	-1.22	.75	.48	---	---	---	---
	(SE)	(.09)	(.09)	(.14)	(.13)	(.11)	(.09)	---	---	---	---
25B. Options are numbers	b	---	---	---	---	---	---	-.23	---	-.24	-.29
	(SE)	---	---	---	---	---	---	(.07)	---	(.07)	(.07)
25C. Options are positive statements	b	---	---	---	---	.53	---	---	---	---	---
	(SE)	---	---	---	---	(.13)	---	---	---	---	---
25E. Options are positive and negative	b	---	---	.86	---	---	---	-.20	---	---	---
	(SE)	---	---	(.34)	---	---	---	(.10)	---	---	---
Multiple R		.64	.61	.65	.62	.46	.41	.27	.16	.42	.35
Cross-validated R		.47	.46	.46	.45	.32	.33	.11	.06	.29	.17

Note. Eq. 1 is based on all variables that added significantly to prediction ($F > 4.0$).
Eq. 2 is based on all variables that added significantly to prediction (increase in $R^2 > .01$).

* Biserials have been multiplied by 10 to show another digit.

Both estimated difficulty and actual difficulty are expressed in delta values.

Table 5

Standardized Regression Weights for the Prediction of Difficulty and Discrimination

Item or Stimulus Characteristic	Item Difficulty (Actual)	Item Discrimination
Usefulness of diagrams (1 = not useful, 5 = useful)	.29	-.20
Number of words in stimulus	.26	-.21
Use of stem "which is a possible sequence, etc."	-.22	.11
Options are lists of names, etc.	-.20	.14
Amount of information to be used (1 = none, 5 = all)	.18	-.12
Number of unvarying assignments	-.13	--
Number of rules or conditions	.13	--
Number of subgroups	--	-.12

Appendix A

Examples of GRE Analytical Reasoning Sets

Appendix A

Examples of GRE Analytical Reasoning Sets

Questions 13-15

The Staff Promotions Committee of a children's hospital must be selected according to the following conditions:

The committee must be made up of four members, two senior staff and two junior staff.
The eligible senior members are pediatricians M and O, dermatologist P, and endocrinologist Q.
The eligible junior members are dermatologist R, who is married to Q; pediatrician T; and dermatologist S.
Doctors M and O dislike each other and refuse to serve together.
No married couple can serve together on the committee.

13. What is the total number of acceptable committees if T becomes ill and cannot be selected?

(A) 1 (B) 2 (C) 3 (D) 4 (E) 5

14. Which of the following conditions, if added to the original conditions, would make P·Q·S·T the only possible committee?

(A) No senior pediatrician can serve.
(B) No pediatrician can serve.
(C) All dermatologists must serve.
(D) Only one representative of any of the medical specialties can serve.
(E) An endocrinologist must serve.

15. Of the following, which is the only committee that is NOT acceptable if P cannot serve and the ban against married couples serving together is lifted?

(A) M·Q·R·S (B) M·Q·R·T (C) M·R·S·T
(D) O·Q·R·T (E) O·Q·S·T

Questions 16-18

The six chemicals manufactured by Apex Laboratories (P, Q, S, X, Y, and Z) are shipped only in Apex's one truck, subject to the following restrictions:

Chemical X cannot be shipped with either chemical Y or chemical Z.
Chemical P can be shipped with any chemical that Apex manufactures.
Chemical S can be shipped with any Apex chemical except Q.
Chemicals Q and Z must be shipped together.

16. Which of the following combinations of chemicals can be shipped together as a complete shipment?

(A) Y, P, S, Z (B) Y, P, Q, Z (C) X, Q, S, Z
(D) S, P, Q, Z (E) X, Q, P, Z

17. How many combinations consisting of chemical X and two other chemicals can Apex ship?

(A) 1 (B) 2 (C) 3 (D) 4 (E) 5

18. Apex ships chemicals X, Y, and S only in 100-gallon containers, and chemicals Z, P, and Q only in 50-gallon containers. What is the total number of different combinations that can make up an Apex shipment if each shipment must contain exactly 150 gallons and each container must hold a different chemical?

(A) 1 (B) 2 (C) 3 (D) 4 (E) 5

Appendix B

Variable Definitions and Instructions to Raters

STIMULUS SITUATION/GENERAL (Ratings That Required Expert Judgment)

1. Usefulness of summarizing the information required for answering items by means of simple drawings, charts, or diagrams

   Instructions: On a scale from 1(not useful) to 5(useful), indicate how useful simple drawings, charts, or diagrams were in helping you to key the items in the set.

2. Number of configurations of entities that are possible under the limitations imposed in the stimulus

   Instructions: If the stimulus is one that uses restrictions to define possible orderings, groupings, or arrangements, generate enough complete orderings, etc., to determine whether the number of all possible ones is ten or fewer or more than ten. If there are ten or fewer, enter the exact number. If there are more than ten, enter 11. If the stimulus is not one that uses restrictions to define possible orderings or groupings, make no entry.

3. Degree to which the problem situation appears to be realistic rather than game- or puzzle-like

   Instructions: On a scale from 1(unrealistic/puzzle-like/artificial) to 5 (realistic), indicate your judgment of the degree of realism in the stimulus material.

4. Relation of the stimulus situation and problem it describes to an identifiable academic discipline

   Instructions: Enter 1 if the stimulus presents a problem or situation in terms that suggest a particular academic discipline or technical field. Otherwise, enter 0.

5. Kind of task to be performed

   Instructions: Enter the code that describes the kind of task that must be performed to solve problems--

   1 ordering

   2 determination of set membership

   3 combination of ordering and determination of set membership

   4 other

STIMULUS SITUATION/GENERAL (Ratings Made by Clerical Staff)

6. Number of sentences in the stimulus (everything between "Questions" and the first item number).

Instructions: Enter the total number of sentences in the stimulus. Count as separate sentences any introductory statements that end with a colon--such as "The following results were obtained:"-- and any sentences that that follow them.

7. Number of words in the stimulus

Instructions: Enter the total number of words (count each letter or number group that is set off from the rest of the text by spacing and/or punctuation as a word--"there are 4 players, L, M, N, O, and P" counts as 10 words).

8. Number of persons, objects, etc., to be ordered, arranged, assigned, grouped

Instructions: Enter the number of persons, objects, etc., that constitute the "pool" from which orderings and arrangements are to be created. (Do not count vehicles, tables, places, etc. that serve as means of grouping or arranging people or things; count only the people or things being grouped or arranged). If the set does not involve ordering, etc., make no entry.

9. Number of subclassifications or subgroupings mentioned

Instructions: If one or more sets of persons, objects, etc., are classified by some characteristic (e.g., sex or color), enter 1. If persons, objects, etc., are classified by more than one characteristic each (e.g., students are identified by both sex and year of graduation), enter 2. Enter 0 if no subgroups are indicated.

10. Number of unvarying assignments of entities to positions

Instructions: Enter the number of cases in which the conditions state that any person, object, etc. is permitted to be assigned to only a single position, situation, or group (e.g., Professor Doe can serve only on committee X; Mary must be scheduled to work on Wednesday).

11. Number of positions in any orderings or groups to be developed

Instructions: Enter the total number of "slots" (positions in a sequence, seats on committees, etc.) that are to be filled. If there is more than one set of slots to be filled, add together the numbers for each set. If the number of slots to be filled is variable or is not specifically stated and cannot be determined by simply counting, do not make any entry.

12. Number of rules, conditions, or restrictions included

Instructions: If the stimulus includes a list of rules, conditions, restrictions, etc., enter the total number. The rules generally, but not always, are introduced by a statement such as ". . .according to the following conditions:," and they are often indented from the text that precedes them.

13. Degree to which the stimulus uses the special language and terminology of mathematics or formal logic

Instructions: Enter the total number of occurrences of phrases from the following list:

"exactly" (followed by quantity)

". . . or . . . but not both"

"if and only if"

14. Number of simultaneous configurations/orderings/ groupings that must be produced using the conditions given (e.g., the number of simultaneous committees or orderings included in a single complete configuration)

Instructions: Enter the number of simultaneous groupings, etc. If no configurations, etc., are produced, make no entry

15. Predominant method of labelling members of the pool of persons or things to be used in orderings and arrangements

Instructions: Enter the code that describes the method used--

1 names of persons or things

2 letters of the alphabet

3 numbers

4 other identifying attributes(e.g., color, size)

16. Composition of the pool of persons or things to be used in orderings and arrangements

Instructions: Enter the code that describes the composition of the pool--

1 living things (persons, animals, imaginary creatures, plants)

2 inanimate objects, events, places

3 other

17. References to objects or people

Instructions: Enter the code that describes references to objects, people, or other beings in the stimulus situation--

1 they are identified by names, numbers, or other symbols (e.g., Mr. Jones, Room 101, a drill press)

2 they are identified only in relational terms (e.g., the third person in line; the uncle of X)

INDIVIDUAL ITEMS (Ratings That Required Expert Judgment)

18. Item classification

Instructions: Enter the code that best describes the item--

1 the item asks what is necessarily true (or must be true)

2 the item asks what can or cannot be true (or what is or is not possible)

19. Amount of information from the conditions that must be used for solution

Instructions: On a scale from 1(none of the information) to 5(all of the information) indicate how much of the information provided in the original conditions must be used to determine the correct answer and/or eliminate the incorrect answers.

20. Perceived difficulty

Instructions: Enter percentage of GRE candidates that you estimate would answer the item correctly.

INDIVIDUAL ITEMS (Ratings Made by Clerical Staff)

21. Use of standard types of stems

Instructions: Enter the the appropriate code for the kind of item stem used (the stems given represent families of stems, and actual stems may vary somewhat in wording).

1 "Which of the following statements must be true?"
(Stem is closed, options are statements)

2 "Which of the following statements could be true?"
(Stem is closed, options are statements)

3 "Which of the following is a complete and accurate list of . . . ?"
(Stem is closed, options are NOT statements)

4 "What is the greatest [or least] number of . . . ?"
(Stem is closed, options are numbers representing quantity)

5 "Which of the following is a possible sequence/ordering/list/schedule/arrangement/etc. . . ?"
(Stem is closed, options are NOT statements)

0 Stem does not belong to any family represented above

22. Whether or not the stem adds new conditions

Instructions: Enter 0 if the stem does not introduce new (i.e., not included in set of conditions that precedes the items) conditions, rules, or constraints. Enter 1 if the stem does add new conditions, etc., to the original ones.

23. Whether or not the stem suspends any original conditions

Instructions: Enter 0 if the stem does not suspend any of the original conditions that precede the items. Enter 1 if any of the original conditions are suspended.

24. Whether or not the stem asks for a negative response

Instructions: If the stem asks for a negative response (all negative response stems include words such as NOT, CANNOT, EXCEPT, FALSE, in upper case as indicated) enter 1. Enter 0 if the stem does not ask for a negative response.

25. Nature of options

Instructions: Enter the code that best describes the options (note that for items that use the Roman numeral format the options are what follows the Roman numerals rather than what follows the letters (A), (B), etc.)--

1 options consist of lists (including lists with only one entry) of names ,numbers, letters, phrases, symbols, etc., that designate beings, places, things, positions, etc. (Options CANNOT be statements.)

2 options are numbers that indicate quantities

3 options are statements, all positive

4 options are statements, all negative

5 options are statements, some positive, some negative

6 options are not of any type described by 1-5 above

26. Use of "can be"/"cannot be" in options

Instructions: Enter 1 if any options use "can be," "cannot be," or similar expressions. Otherwise, enter 0.

27. Use of "must"/ "must be" in options

Instructions: Enter 1 if any options use "must," "must be," or similar expressions. Otherwise, enter 0.