RR-87-44

RESEARCH REPORT

A THEORETICAL FRAMEWORK FOR THE STUDY OF ITEM DIFFICULTY AND DISCRIMINATION

Janice Dowd Scheuneman
Karin S. Steinhaus

Educational Testing Service
Princeton, New Jersey
December 1987

A Theoretical Framework for the Study of Item Difficulty and Discrimination

Janice Dowd Scheuneman

Karin S. Steinhaus

Educational Testing Service

Princeton, New Jersey

A Theoretical Framework for the Study of Item Difficulty and Discrimination

Janice Lowd Scheuneman

Karin S. Steinhaus

Educational Testing Service

ABSTRACT

Traditionally, test item difficulty is a statistical concept, defined in terms of the performance of examinees rather than in terms of intrinsic properties of the item itself. A clearer understanding of the association between item properties and examinee performance, however, would result in numerous benefits, including better prediction and control of item difficulty in the test development process and enhanced construct validity of the test. As a first step toward the goal of achieving such understanding, a theoretical framework is delineated, drawing on both measurement concepts and concepts drawn from cognitive psychology and personality theory. In this formulation, the difficulty of an item is seen as a function of the demands set by the item tasks and the abilities and attributes which the examinee may find necessary or useful in responding correctly to the item. In addition, interactions of examinee abilities and item characteristics may occur where solutions to an item may be reached by using different strategies or abilities and the difficulty of meeting the item demand using these different approaches is not equivalent. Examples from the measurement and psychological literature provide suggestions of a number of examinee characteristics and item properties which might be expected to affect item difficulty. The formulation is then extended to item discrimination. Finally, the literature on verbal analogies is reviewed within the theoretical framework to suggest sources of variation in the difficulty and discrimination of this item type.

A Theoretical Framework for the Study of Item Difficulty and Discrimination

Janice Dowd Scheuneman, Karin S. Steinhaus

Educational Testing Service

In a recent papers, Glaser has pointed out that psychometric research has focused almost exclusively on the end product of testing--the data resulting from persons' responses to test items--while little systematic study has been made of the preparation of tests and test items at the front end of the process (Glaser, 1986; Glaser & Lesgold, 1985). After nearly 70 years of objective group measurement, item construction remains largely an art form or a skilled craft practiced by those with an aptitude for this task. At present, even experienced practitioners of this art have been found to be unable to estimate accurately the difficulty of items for a population with which they were familiar (Bejar, 1983). Despite their skill in preparing items which function properly for a specified purpose, they have probably had little awareness of many of the possible sources of variation in item difficulty. Measurement research has provided little guidance in this task

In the psychometric tradition, item difficulty has been defined in terms of the performance of examinees rather than in terms of any intrinsic properties of the item. In classical test theory, difficulty is defined in terms of the proportion of examinees producing a correct response to the item. Modern theory has freed the definition of difficulty from the characteristics of a particular sample of examinees, but the difficulty parameter is still defined in reference to examinee ability levels. If no guessing occurs, the ditficulty of an item is the level of ability on the unidimensional trait measured by the item at which the probability of a correct response is .50.

Both modern and classical measurement models have more than proved their worth in representing item functioning in a wide variety of practical measurement problems and applications, but as a source of explanatory principles of item difficulty, both theories are inadequate in their present state of development.

Intrinsic item difficulty, on the other hand, would be defined in terms of the item content, context, characteristics or properties and the task demands set by the item which must be met by an examinee with an assortment of skills and abilities in order to produce a correct response. To the extent that we can come to understand more fully the intrinsic difficulty of items, we can also begin to understand better the functioning of test items and to bring that functioning increasingly under control. A number of benefits might then accrue, including: (a) fewer items lost in pretest, (b) better control over test properties in programs not pretesting, (c) more precisely delineated content specifications, (d) better diagnostic information, (e) improved quality of judgments for standard setting procedures, (f) more rational defense of individual items where challenges occur, (g) enhancement of knowledge base required to make feasible the computer generation of certain types of test items, and (h) improved construct validity. If the development of test items is to move from art or craft toward science, a theory is needed which would permit the generation of testable hypotheses concerning the major relevant components of intrinsic item difficulty.

In this report, a conceptual model designed to consider a variety of sources of item difficulty will be developed. This model provides the framework for a theory which is broad enough to encompass contributions to difficulty made as a result of different component tasks of a test item, both

cognitive and noncognitive, and the interactions of the task demands with various characteristics of the examinees. The model is intended to (a) elicit ideas about difficulty and discrimination which might not arise under a more limited conceptualization; (b) provide a framework broad enough to encompass and integrate a full program of research, as well as to provide direction for that work; and (c) to provide connections between and among research results which may not be readily perceived to relate to issues of difficulty of test items. As the discussion proceeds, references to the relevant literature will be included. The last section of the report provides an example of the application of the model with respect to verbal analogy items.

The Model

The concept of "difficulty" is largely meaningful in reference to measures of cognitive abilities where one response to an item is considered "correct" and the others "incorrect." In measures of attitude, interest, or personality, for example, it may be reasonable to consider the probability that a given response will be selected by examinees with a particular attribute of interest, but this response likelihood is unlikely to be referred to as "difficulty." In these instances, although different responses to an item may be construed as reflecting different attributes or levels of an attribute, no one particular response to an item is likely to be considered "correct." One could, no doubt fruitfully, consider the various sources of response probability in these types of instruments, but the discussion here will be limited to tests intended to measure cognitive abilities and the "difficulty" of the items appearing on such tests. The model will, however, include the effects of non-cognitive variables on examinee performance on the cognitive tests and hence on their item difficulty.

For each cognitive test, a particular ability or constellation of abilities exists that the test is designed to measure and that is generally specified by the purpose of the test. For example, the intended ability may be competence in some achievement domain, such as an area of school instruction, or the skills, knowledges and abilities required for competence in an occupation or profession, or the ability to be measured may be general, such as spatial or perceptual ability or aptitude for music or for learning in a college instructional setting.

Although the model for item difficulty is stated below in terms of equations, these equations should be considered heuristic devices specifying classes of variables rather than mathematical formulations. A number of mathematical models of difficulty for narrowly defined item types have been developed which could be subsumed as special cases within the framework presented here. (See Embretson, 1983, for a review of some of these models.)

Many item response theory (IRT) models require that the intended ability be unidimensional, but even in tests where unidimensionality has been satisfactorily demonstrated, this ability is usually defined to have several facets, each of which may be considered important and sufficiently discrete to justify specific inclusion in the test construction plan. Hence, the unidimensionality requirement is assumed to be met if the items function as if they measured a single underlying trait, rather than that they conform strictly to a single dimension, as would be the case with a pure Guttman scale where a person getting one item correct may be assumed to get all easier items correct. The model being developed here assumes that a number of different abilities or attributes will affect the response to different items, but the resultant test score may still be "unidimensional" in the sense that it meets

the quantitative criteria usually applied for assessing dimensionality prior to the implementation of IRT procedures.

The standard definitions of item difficulty based on examinee performance can be restated in a form somewhat different from the usual for the purposes of this argument. Let the observed item difficulty in some metric (from either classical or modern theory) be defined as follows:

$$D_i = \Theta_g + \delta_i + \varepsilon_{gi} \qquad (1)$$

where

D_i - observed difficulty of item i

Θ_g - the true ability of examinees in group g on the trait the test is intended to measure

δ_i - the level of that ability demanded for the task set by item i, and

ε_{gi} - error

In classical measurement, D_i might be the proportion of correct responses in a sample; Θ_g, the mean test score of the sample or the expected proportion correct across all items of the test; and δ_i, the extent to which the ability demanded by the item is greater or less than the mean ability of the sample. The D_i in modern theory (often referred to as item response theory or latent trait theory) would be the estimated value of the difficulty parameter, b. Θ_g would be a scaling constant or a reference value on the theta (ability) scale, and δ_i the true value of difficulty on the theta scale. For example, if Θ_g were set to zero, the parameter estimate b would equal the true parameter plus or minus the error of estimation. In both instances, the variation in item difficulty in a given test as administered to a given sample or in relation to a given reference value is a function of the different levels of ability (δ_i) demanded by the item.

In a real life testing situation, however, an examinee will not typically restrict herself or himself to the use of those abilities the test is intended to measure, but will bring to the testing task a whole constellation of other abilities, attitudes, values, and personality traits which will affect her or his response to the item. We may then improve our model by defining the difficulty of an item in terms of the demand placed on any or all of the abilities and attributes which may be required, or are merely useful, in responding correctly to the item, including but not restricted to the ability the item is intended to measure. If we wish to think of intrinsic difficulty, rather than observed difficulty, the various components of difficulty may be considered without regard to the levels of those abilities or attributes in any particular examinee group. For example, an item may be either too easy for a group (most or all of the examinees are able to get the item correct), or too difficult (few if any of the examinees have the skills required to respond correctly to the item).

In turn, items may differ in the degree to which any particular ability or trait will be useful in meeting its demands. For some items, the task demands may be met only through the use of a single ability, with success on the item dependent solely on whether the level of ability demanded is within the capability of an examinee. For other items, the demands might be met in a variety of ways so that different abilities may be used in arriving at a correct response or one ability may be substituted or combined with another. In these instances, some examinees may be more adept than others at selecting which abilities will be most useful in responding to the demands of an item. For other items, lack of sufficient levels of the intended abilities might be compensated for by using other abilities with which the examinee is more

skilled or knowledgeable.

A first step in improving our item difficulty model, therefore, is to include all the abilities which an individual might use in meeting the task demands of an item. With the addition of a term to represent the abilities that the test was not explicitly intended to measure, the formulation of the model becomes

$$D_{ig} = \Theta_g + \phi_g + \delta_i + \varepsilon_{gi} \qquad (2)$$

where D_{ig} is the difficulty of item i for group g, Θ_g the level of the examinees' intended ability in group g, ϕ_g other abilities and attributes that may be used by individual examinees in group g in meeting the task demand, δ_i the demand of the item on these different abilities and attributes, including both Θ_g and ϕ_g, and ε_{gi} is again the error. Note that, since the item demands represented by δ_i may be met by different examinees using different abilities, the demand is not attached to a particular ability in the model.

Clearly, this model suggests that a very large number of skills, abilities, knowledges, and attributes are relevant to the intrinsic difficulty of an item. Depending on the purposes for which the item is being studied, however, different aspects of difficulty will become important. If the reason for studying difficulty is to gain better understanding and control of the observed difficulty of items in actual tests when administered to specified examinee populations, it is possible to impose limitations on this set of variables. Variation in the observed difficulty among the items in a given test will not result from all possible demands set by each item nor all possible abilities which might be brought to the item task by an examinee. Criteria can therefore be established to delimit a subset of item

properties and examinee abilities and attributes that are most likely to be of interest in this context.

First, an item demand for a particular ability will contribute to the overall level (mean) of observed item difficulty on a given test if and only if the skills needed to meet that demand are beyond the capabilities of some of the examinees. That is, if a given task demand can be met by all examinees on all items, no decrement in the probability of a correct response due to examinee deficiencies in that ability can occur. For example, the ability to encode words written in English using the standard Latin alphabet is not likely to effect the difficulty of items for American college students without visual impairment whose first language is English, even though this skill may in fact be demanded by all the items on a particular test. For a group of these same college students who are beginning the study of Arabic, however, encoding a word written in Arabic script may indeed contribute to the observed item difficulty in an Arabic vocabulary test.

If the task demand of interest is beyond the capabilities of some of the examinees, it will still not contribute to the variation in observed difficulty among the items of the test unless the items also differ in the degree to which they demand that ability. For example, encoding ability may contribute to variability in item difficulty in a written vocabulary test for beginning learners in a language such as Greek or Russian which use some letters unlike those in the Latin alphabet and other letters which look alike, but correspond to different sounds, as well as letters which are the same in the different alphabets. Not only may students differ in the extent to which they have mastered the encoding task, but the words may differ in the extent to which they demand the encoding ability depending on the particular letters

composing a word. That is, a vocabulary word which is composed entirely of letters which are the same as in English will present an easier task than a word composed entirely of letters unlike the more familiar forms.

Notice that if the Russian test in the above example was intended only to measure the students' skill in transliterating the Russian words into the more familiar Latin alphabet, variation in the ability to encode within a given sample of examinees would produce variation in test scores, but would be insufficient to produce additional variation in item difficulty. That is, the difficulty of the items in this test would vary only with regard to the difficulty of the encoding task, although the mean level of difficulty would be a function of the mean level of the intended ability of the examinees as in equation (1) above.

To the extent that other abilities are required, however, individual variations in abilities will also contribute to observed difficulty. Suppose, for example, that the test in beginning Russian is intended to measure Russian vocabulary knowledge at an appropriate level. Some examinees may recognize and know the meaning of a word if spoken, but are unable to encode it when it is presented in its written form, while others may be able to encode and pronounce a word, but fail to know its meaning. That is, in instances where more than one ability is required to respond correctly to an item, or when an item demand may be met using more than one set or combination of abilities, and the other criteria mentioned above are also met, individual differences in abilities will contribute to the variation in observed item difficulty.

For example, think of equation (2) above as a regression equation for predicting the difficulty of the items in a test for a particular examinee population. Notice that in this context, θ_g might be seen as a regression

constant. This may be clearer in an IRT conceptualization where θ is only a reference value. For a given test and a given sample, however, it may also be treated as a constant in classical terms. That is, the level of the ability the test is intended to measure may be treated as constant for that group on any given occasion, which might, for example, be represented by the mean difficulty of the items in the test. The term ϕ_g then represents the independent contribution to successful performance on the item of individual abilities and attributes other than the intended ability and λ_i is the demand of item i for each of the different abilities and attributes in both θ_g and ϕ_g.

Providing that all of the above criteria have been met, another criterion for the inclusion of a particular attribute or ability in the model of observed item difficulty is that it have a reasonable probability of being used. That is, if an item <u>can</u> be solved using a particular subset of abilities and attributes, but is very unlikely to be solved in this way, those attributes and abilities are less likely to be of interest. If a researcher were developing an elaborated model of the type presented here for a specific item type or testing instrument, a term for the probability of use of the various abilities and attributes might therefore also be included.

One last criterion for an item demand for an examinee ability or attribute to be of interest in contributing to the variability in observed item difficulty is that the ability or attribute not be highly correlated with the ability the test is intended to measure. Since the test items are generally constructed to place a demand on the intended ability, the item task demand is likely to be greater for that ability than for any of the other abilities or attributes that may used in making a correct item response.

Hence, to the extent that an attribute covaries positively with the intended ability, the demand on that attribute is likely to be met if the demand on the intended ability is met. Conversely, when the correlation between some attribute and the intended ability is low, the more likely it becomes that the demand on one can be met while that on the other cannot. This leads to the interesting speculation that some of the attributes and abilities that are important in predicting observed item difficulty may be those with relatively low correlations with the ability the test is intended to measure.

In summary, intrinsic item difficulty is a function of the demands set by the item task and all abilities or attributes which may be used to meet those demands. Observed item difficulty, on the other hand, will be concerned primarily with those abilities and attributes (a) that are beyond the capabilities or outside the propensities of some examinees in the population of interest, (b) for which items differ in their demand, and (c) for which the individual capabilities and propensities vary. Further, the abilities and attributes most likely to be of interest are those that have some probability of being used and that have relatively low correlations with the intended ability.

Individual Difference Variables

For most tests of academic aptitude and achievement, the abilities or attributes most likely to result in variation in observed item difficulty will be cognitive abilities. Non-cognitive variables will also influence the difficulty of items, however, both directly and indirectly through their effects on cognitive functioning. In the following sections, the literature concerning the psychological components of test performance is reviewed to identify possible contributors to item difficulty which may stem from examinee

abilities and attributes.

Consistent with the criteria stated in the discussion of the model above, this review focuses on cognitive and non-cognitive variables which show individual differences. The effects of some of these variable may not be felt in the observed difficulty of the items on a particular test, however, or may be felt only in the overall level of difficulty rather than in the variation in difficulty among the test items. That is, for a given test and a given examinee population, it may be that either the demand on an ability set by any of the items is within the capabilities of all examinees (no effect) or is beyond the capabilities of some examinees, but the demand for the ability does not vary over the items of the test (only the overall difficulty level is affected). This possibility was not considered in the following summaries, where any potentially relevant individual difference variable identified from the literature is briefly discussed.

Cognitive Process Variables

Within cognitive psychology a body of literature is developing which describes the component cognitive processes used in solving the tasks set by test items or other stimuli resembling test items. Most of this work, however, has concentrated on processes which are used by all subjects or examinees, making it of little interest in the present context. Nevertheless, some suggestions for individual difference variables which might affect examinee performance and hence item difficulty can be found.

Carroll (1976) uses an analogy to computer information processing in discussing individual differences in "production systems." A production system is a set of condition/action statements or rules concerning actions to be taken given certain conditions. Individual differences may arise in

(a) the composition and ordering of the condition/action rules incorporated into the system, (b) the particular action strategies used and (c) the data available to the system. Other differences may arise in temporal parameters or in the success of the individual in applying these rules.

Snow (1980) also discussed individual differences in processing rules as they are applied in the test taking situation. He suggested that performance differences arise from individual differences in (a) the efficiency of organizing processing strategies, (b) the control one exerts over this organization and (c) the ability for sustained application of these rules throughout the entire test.

Similar abilities were discussed by Sternberg (1985) in his triarchic theory of intelligence. In this theory, he proposed three subtheories concerning the functioning of intelligence in the test-taking situation of which two are pertinent here. The first of these is the subtheory addressing different information processing components which include "metacomponents," higher level executive processes. Individual differences in any of the following metacomponents may contribute to differences in item performance: (a) deciding on the nature of the problem to be solved, (b) deciding on the performance components relevant for solving the item task, (c) deciding on how strategically to combine performance components, (d) selecting a mental representation for information, (e) allocating resources such as time for problem solution, and (f) monitoring solution processes. The second subtheory concerns the previous experience of the examinee with the tasks or situations presented by the item. In particular, this subtheory concerns the degree of novelty of the task for different examinees and, conversely, the degree to which performance has been automatized prior to the examinee's taking the

test

In tests of achievement, one may improve performance by generalizing from a subject matter area one knows very well to an area one knows less well. Messick (1984) has pointed out that as a person learns a field, he or she develops strategies for acquiring, structuring and retrieving information. In a testing situation, an examinee may be able to make use of strategies learned in studying a field not being measured to generate hypothesized information beyond that provided in the item. These hypotheses may then be used in reaching a correct solution even when the requisite knowledge is not present (i.e., the task demand for the intended ability is beyond the capability of the examinee).

Non-Cognitive Variables

In addition to the cognitive process variables is a host of attitudinal and personality variables which may also affect performance. In a recent paper, Messick (1985) discussed personality traits and styles which might be expected to influence cognitive functioning or performance on cognitive tasks. Some of these are not strictly separate from cognitive variables; Messick has called them "ability-personality admixtures." Some of the variables he discussed are independence, carefulness, self-assurance, self-control, criticalness, rigidity, alertness, impulsivity, tempo, energy expenditure, self-sufficiency, tolerance for ambiguity, inhibition, ability to mobilize, surgency, confidence, suspicion, stability, and endurance. Any of these may be expected to influence performance on the cognitive tasks required in tests of aptitude or achievement under at least some circumstances. He also mentioned personality traits explicitly related to measurement, including the propensity to guess, use of partial information, tolerance for different types

of errors, risk taking, evaluation anxiety, and impression management.

Not all of the abilities or personality characteristics which produce individual differences, however, are functionally operative for all individuals to the same level of proficiency or intensity, nor need all these dimensions even be present in all individuals. A critical source of personality differences derives from precisely which traits are central, important, or valued by the person. Moreover, persons will differ in relative trait level or intensity and relative strengths and weaknesses of various traits, as well as within person patterns of trait interrelationships. An example of such patterns would be cognitive styles, defined by Messick as "characteristic self-consistencies in information processing that develop in congenial ways around underlying personality trends" (1985, p. 36).

Another source of individual differences is the affect experienced by the examinee surrounding both the learning and the testing situation, one obvious example being test anxiety. Each person's past history will, of course, determine which particular situations, contexts, or other stimuli will produce positive or negative affect. The ebb and flow of the individual's investment of affect in ideas and ideologies, his or her interests and other intrinsic motivators will influence the relative salience or strength of different traits. As Messick (1985) points out, how positively individuals learn to feel about themselves and others, as well as about different subject-matter fields and ideologies, shapes the development of their knowledge and ability structures with implications for preferred methods of inquiry and ways of knowing, as well as for the content of things known.

Components of Item Task Demand

Which of the various abilities and attributes will be brought to bear

by the examinee in solving a test item depends, of course, on the demands of the item. The item demands may be created through a number of mechanisms which may be classified into the following categories: (a) manifest content, which sets both the knowledge and process requirements; (b) item properties, such as the format of the item, which may serve to mediate how well the manifest content requirements are apprehended; and (c) characteristics of the test or the context set by other items on the test, which may affect the examinee's perceptions or expectations concerning the item task.

Manifest Content Requirements

In an achievement test, the main outlines of the demands to be placed on knowledge of the intended achievement domain are explicitly set forth, at least in part, in the test specifications. Items are developed to measure specific facts or concepts from the knowledge domain at a level generally appropriate for the intended test use. Test constructors can generally control difficulty to some extent through manipulation of the level of knowledge required by the item. In addition, however, incidental demands on knowledge are part of items in both achievement and aptitude tests. In achievement tests in many areas of science, for example, a certain level of skill in mathematics is also required. In many mathematics tests, a certain verbal facility is needed in order to understand the nature of the problem to be solved. Many of the tests of scholastic aptitude or intelligence require a basic level of language skills, some knowledge of mathematics, and a number of commonly known facts. In many instances, these incidental demands are assumed to be unimportant, and rightfully so, since they are well within the capabilities of all examinees in the population to be tested. This assumption is not always correct, however; the facts may not after all be known or

vocabulary may not be recognized. These incidental demands may then indeed provide a source of variation in item difficulty.

Process requirements of an item might best be stated in broad terms as the item demands might be met using a variety of cognitive processes and strategies. One commonly used schema for describing process demands is Bloom's taxonomy, which includes knowledge, comprehension, application, analysis, evaluation, and synthesis (Bloom et al, 1956). Messick (1984) has suggested the following list of process requirements which are tied more closely to current research in cognitive psychology: comprehension, retrieval from memory, visualization, restructuring, reasoning, and judgment.

More recently, Emmerich (1986) developed a classification scheme that includes both knowledge and cognitive demand components and takes account of the more recent research findings. His cognitive demand categories elaborate on Bloom's taxonomy and include five major divisions, each of which has some small number of subdivisions. The major divisions are synthesize, support or weaken, analyze, identify, and restate. Emmerich's second taxonomy concerns aspects of knowledge and includes six major categories, three of which have subdivisions. These include language, entities, relationships, procedures, criteria, and theory.

Research on Bloom's taxonomy, however, has failed to demonstrate a clear link between these process variables and item difficulty or other properties which might be expected to relate to difficulty (Blumberg, Alschuler, & Rezmovic, 1982; Seddon, 1978). The more recent conceptualizations of process variables may produce better results, but it seems likely that the effects of the process variables are confounded with other sources of difficulty. Multivariate designs may therefore be necessary if a link between process

requirements and item difficulty is to be demonstrated.

Item Properties

Some characteristics of items which may affect difficulty and discrimination include (a) the format or structure in which the item task is presented; (b) the mode of presentation, such as verbal, numerical, or figural modes, which may be used, for example, for different items in a math test or in a group-administered intelligence test; (c) the number and difficulty of words and semantic properties of items containing verbal material; (d) the use of symbols, charts, or diagrams in various types of achievement test items; and (e) various properties of figural stimuli used in spatial perception items.

Much of the research on the effects of item properties on difficulty and discrimination has focused on characteristics of format. Dudycha and Carpenter (1973) observed that open-stem items were more difficult than closed-stem items. (In closed stem items, the question is a complete sentence. In open stem, the options complete a sentence begun in the stem.) They concluded that item difficulty can be changed by altering either the openness of the item stem or its positive/negative orientation (but not both) without adversely affecting its discriminatory power. They also found that inclusive options ("all of the above") significantly decreased the discriminability of an item. Hughes and Trimble (1965) found that complex options ("all of the above," "none of the above," "both 1 and 2 are correct") increased item difficulty but had inconsistent effects on discriminating power. Williams and Ebel (1957) found that in 2- to 4-choice items, decreasing the number of choices decreased the difficulty considerably and the discrimination somewhat, although two-choice items were much more quickly answered than four-choice

items.

Forsyth and Spratt (1980) investigated multiple-choice math items with variations in item format to introduce one-step and two-step operations necessary to the solution of the items. They found that the two-step format tended to produce more difficult items and to lower the discriminating power of the item, but doubted that the two formats measured the same construct. Owens, Hanna, and Coppedge (1970) studied the effects on the difficulty of geometry items of judgmental factors (plausible distractors), error frequency (using typical student errors as distractors), and discrimination (using discriminating errors as distractors). They found that the three test versions were equally valid, with no differences in discriminability, but the reliability of the judgmental tests was inferior when compared to the other two tests.

In these studies, the difficulty and discrimination of two or more sets of items with different format characteristics was compared. These studies may therefore be criticized because the effects on item difficulty or other test properties may have resulted because the different formats measured somewhat different constructs. More recent studies have used multivariate designs to predict item difficulty from various item properties within an item set.

Stenner, Smith, and Burdick (1983) developed a theory of receptive vocabulary which hypothesized a number of specific relationships between item difficulty and some characteristics of the words used in items of the Peabody Picture Vocabulary Test. They were able to predict approximately 70 percent of the variance in item difficulty from these vocabulary variables. Smith and Green (1985) were also able to predict the difficulty of items on a

paper-folding test from various features of the stimulus. Embretson (1985) evaluated a number of different models of prose complexity to account for the variation in difficulty of paragraph comprehension items. Similarly, Bejar and Yocum (1986) were able to model difficulty in hidden figures items.

Test Characteristics

Although the focus in this paper is on individual items rather than the test as a whole, the difficulty of items can also be influenced by the context set by the total test. One such test characteristic that might be said to affect item difficulty is the adequacy of the instructions or task "set," that is, the general task requirements of all items or of a recognizable subset of items. If instructions are ambiguous, examinees may differ in the degree to which they understand the task to be performed. Further, an imperfect understanding may lead to a correct solution to some items and not to others. For example, if a child does not understand what the task is that is set by analogy items, she or he may infer that the correct approach is to find in the list of options a pair of words which are synonyms or antonyms. This strategy will lead to a correct response where this is indeed the relationship required by the analogy item and an incorrect response when it is not. That is, the child may fail to understand that the stimulus pair serves to identify the appropriate relationship for a given analogy item.

For young children or for persons from cultures where they have had little previous experience with testing, even the mode of expressing their response may be a source of difficulty. For example, learning to handle an answer sheet or to fill in bubbles with a No. 2 pencil may distract attention from the testing task. Test length and time limits may also affect the difficulty of items which are near the end of a test, either through fatigue,

insufficient time to adequately consider the items, or failure to reach them at all Conversely, within-test learning may result in the early items in the test being relatively more difficult than those that occur later.

Item difficulty has also been shown to differ according to the context set by other items, including possible order effects and the content and average difficulty of the other items or the test. The literature on such effects has recently been reviewed by Leary and Dorans (1985). Below is a brief synopsis of the findings they report.

Studies on item order have most frequently involved comparing performance on groups of items that have been assembled into test forms in which easy, medium and difficult items appear in varying patterns. The effects of changing the order of difficulty have generally been non-significant. Where differences have been found, the easy-to-difficult sequence appears to result in higher scores. The apparent superiority of this sequence may, however, be explained by the effects of speededness. If relatively easy items appear near the end of the test, candidates may not reach them before the test is over. Under strictly power (or near power) conditions significant results were found only for aptitude or mathematics achievement tests. In verbal aptitude tests, items that appeared late in the test were found to be more difficult than the same items appearing early in the test when the easy-to-hard sequence was held constant for the other items. Whitely and Dawis (1976) obtained a similar result. They determined that the sequencing of verbal analogy items can significantly influence the difficulty levels of the individual items.

Other studies have investigated the interaction of item order with test anxiety, sex, and levels of achievement. Generally, test anxiety has not been found to interact with item order. One study was found which showed an

interaction with types of anxiety, but it was not replicated. The order of items in math tests has been found to have different effects on performance, depending on the sex of the candidates.

In tests where similar items are grouped together in sections, the placement of the section in the test may also affect the difficulty of the individual items within each section. For example, if items of a certain type appear in a section that is placed after a section consisting of similar items, the difficulty of these items may be affected either by within-test practice effects or fatigue. Leary and Dorans (1985) reported on several studies that found at least some items that showed such within-test effects.

Interaction of Individual Differences and Task Demand

The model as it has been posited to this point states that item difficulty is predicted from the demands of the item task and the capabilities of the examinees in meeting those demands. These capabilities are assumed to include both the abilities that the test is intended to measure and whatever other abilities or traits may be required in reaching the correct solution to an item. In many cases, however, particularly as items become more complex, the item demands may be met in a number of different ways so that different abilities may be brought to bear on the item by different examinees.

In instances where the task demand can be met in more than one way, the possibility for an interaction between examinee abilities and the item demand exists. Certain conditions must be met, however, for such an interaction to influence observed item difficulty. In order to describe these conditions more clearly let us assume that a limited set of approaches to meeting the task demand are available and that a probability that a given approach will be used by some subset of examinees may be attached to each of these alternative

approaches. An interaction requires that these probabilities differ for subgroups of examinees defined by their capabilities or patterns of capabilities on the various aptitudes, traits, skills, and knowledges measured by the test or by attitudes, values, or personality characteristics associated with performance on the items Further, an interaction requires that the difficulty of meeting the task demand is not the same for those alternative approaches which are most likely to be used by the different subgroups. That is, the difference in approach to a problem used by different groups is important primarily if the approaches are not equally likely to produce a correct response.

The result of this interaction will be that the item difficulty can be observed at one valu with one subset of examinees and another value with a different subset. The item difficulty model may thus be expanded to include an interaction term, as follows:

$$D_{ig} = \Theta_g + \phi_g + \delta_i + \phi_g \delta_i + \varepsilon_{ig} \qquad (3)$$

where $\phi_g \delta_i$ is the interaction and the other terms are defined as in Equation (2).

In a slightly different context, Snow and Peterson (1985) gave examples of this type of interaction. Their first example was from a study by Gavurin (1967) in which the test task was the solution of anagrams. In one condition examinees were free to manipulate tiles on which letters were written; in the other condition they were not. In the former condition, difficulty as measured by time to solution was negatively related to spatial ability; in the latter condition the difficulty was positively correlated with spatial ability. Hence, where mental manipulation and visualization of the letters was a useful strategy, spatial ability enhanced performance; where this task

demand for spatial ability had been removed, it had a mildly decremental effect.

The second example given by Snow and Peterson (1985) was from a study by Schmitt and Crocker (1981). In this study, the first condition was one in which examinees were required to generate their own response to an item before viewing the options. The second condition was the standard multiple-choice item format. The individual difference variable was a test anxiety score. It was found that examinees low in anxiety performed better in the condition where they formed their own response first, but high anxiety examinees did better in the standard condition.

Chaffin and Pierce (1987) provide an example of item difficulty interactions with analogy items from the Graduate Record Examinations General Test. The analogies were classified according to the analogical relationship. These relationships were then further categorized as conceptual (happy/sad, large/small) or pragmatic, that is, defined by usage (tailor/sew, physician/patient). After controlling for verbal ability, students from fields such as engineering, math, or computer science were found to do better on the items with pragmatic relationships while students from more verbal areas, such as English and history, did better on the items with conceptual relationships.

A different kind of interaction variable is the individual examinee's test-wiseness skill. Test wiseness differs from other kinds of abilities in that it can be used to enable the examinee to replace the intended task demand with another so that the abilities required are different than for other examinees who either do not possess or do not choose to use test-wiseness skills. In such instances it is possible to produce a correct response

without the requisite levels of the intended abilities set by the task demand. To the extent that an item is susceptible to these strategies, an interaction may appear with the item showing one level of difficulty for those with the requisite test-wiseness skills and another for those who are meeting the task demands in the intended fashion.

Studies have been done to examine the effects on item difficulty of test-wiseness cues, such as those articulated by Millman, Bishop, and Ebel (1965). In general, these studies have contrasted the difficulty and discrimination of items with and without test-wiseness cues without consideration for individual differences in either the ability or the propensity to make use of these cues. Interactions have been found with other individual difference characteristics, but in many studies failure to take account of these individual differences may have resulted in a lack of significant results or small effect sizes. Nonetheless, results give some indications of the contribution of test wiseness to item difficulty and suggest some areas where interactions may exist.

In an investigation of performance differences of Black and White examinees, who might be assumed to differ on a number of abilities and attributes, Scheuneman (1987) found an interaction of group membership and test-wiseness cues in specially prepared items in a verbal section of the Graduate Record Examination General Test. The items were constructed in pairs that differed only in the presence or absence of test-wiseness cues in the options. The difference between the difficulty of the paired items was larger for White than for Black examinees. Combined with other findings from the study, this led the author to hypothesize that Black and White examinees differed in the kinds of test-wiseness cues that were used rather than in

whether or not these cues were used at all.

Board and Whitney (1972) also found an interaction when they tested undergraduate students in a course in American Politics. The inclusion of extraneous material made items easier for poor students but more difficult for better students, a result that also had the effect of reducing the internal consistency of the test. They also found that (a) incomplete stems made the test items more difficult and reduced internal consistency; (b) when the key was a different length from the distractors, the test items were less difficult for poor students, reducing internal consistency and validity of the test; and (c) grammatical consistency between stem and keyed response did not have a major effect on item difficulty. They concluded that poorly written items, which are often those most susceptible to test-wiseness strategies, obscured differences between good and poor students.

Plake and Huntley (1984) examined internal context cues provided by singular/plural forms and by initial vowels and consonants that may serve to indicate the key in certain item types. They found such cueing effects to be minimal. Dunn and Goldstein (1959) found that items containing internal cues to keys, extra-long keys, and inconsistencies in grammar between key and distractors were found to be less difficult than items written according to standard test development rules. The researchers found no significant effect on reliability or validity that could be attributed to violation of any of the rules. Similar results were found by McMorris, Brown, Snyder, and Pruzek (1972).

Strang (1977) examined non-content cueing due to option length and level of language technicality. He found that long non-technical options were chosen more often than other types of options. Green (1984) varied items by

increasing stem length and syntactic complexity and by substituting uncommon terms for more familiar terms in the stem. The effects were unpredictable; increasing language difficulty appeared to add information but to make items either easier or trickier as a consequence. Green also varied semantic similarities in options to create three levels of option convergence. These results indicated significant effects on difficulty.

Measurement Error

In the measurement models proposed by both modern and classical measurement theory, the error term combines the normal sources of measurement error with any of the sources of variation in item difficulty other than those which are associated with the ability the test is intended to measure. The model given in Equation (3) could, theoretically, represent all the systematic sources of variation in item difficulty that are properties of either examinees or items. The error term then would represent only random effects or unstable conditions associated with a particular administration. These might include sources of measurement error affecting all examinees such as serious distractions during testing, adverse conditions in the testing room (the air conditioning has broken down), or those affecting only some examinees such as temporary memory retrieval difficulties or various indispositions (a number have colds or smokers are not permitted to smoke in the testing room).

Item Discrimination

Although item discrimination is mentioned occasionally in the above discussion, the model presented in Equation (3) is stated only for item difficulty. Discrimination is similarly supposed to be affected by the item demand on both intended and incidentally measured abilities and attributes.

Even in theory, however, item discrimination seems likely to require a more complex representation than that suggested by a linear combination of effects and their interactions. Nonetheless, some generalities may be stated within the framework developed here.

In item response theory, the discrimination of an item is represented by the slope of the item characteristic curve; that is, for a highly discriminating item, the probability of a correct response rises more rapidly as ability increases than for an item with lower discrimination. If the task demand of an item were limited entirely to some level of the ability the test is intended to measure, the probability of a correct response might be expected to remain at zero until the requisite ability level is reached and then to become one. The discrimination should be perfect (or nearly perfect with some slight allowance for measurement error); the biserial correlation should be one. This would be the case with a perfect Guttman scale. It is hypothesized here that, as incidental abilities (expressed in the model above as :) are demanded by the item task or can be used by an examinee to meet the item demand, the observed item discrimination is reduced.

The ability the test is intended to measure may not be unidimensional in this strict sense, but as the ability becomes more complex, the discrimination would theoretically decrease. Further decrements might then be expected as unintended abilities or attributes also affect performance on the item. For example, test developers often encounter difficulty in producing items that are hard for able examinees that also have adequately high biserial correlations. Such items may be difficult because they demand a high level of some of the incidental abilities rather than demanding a high level of the intended ability. A low discrimination index may thus be an indicator that

other abilities are being demanded or that an interaction exists affecting a substantial proportion of examinees.

In terms of construct validity, however, the ability that the test is intended to measure may not be the only important ability for the purposes of the test. If the construct we are attempting to measure is multidimensional and our instrument taps only some of these dimensions well, some of the abilities or attributes which appear to be incidental, may in fact be valid with regard to the purpose of the test, even though they were not "intended." Perfect discrimination of the type which would result from a strictly unitary ability is, therefore, not necessarily desirable. The development of an understanding of the abilities and attributes contributing to item difficulty in a given test may also allow us to determine if these factors result in a decrement in the item's discrimination and, if so, whether or not the measurement of these factors is desirable.

Verbal Analogies: An Example

As stated earlier, the theory which has been developed here and its associated models are intended primarily as heuristic devises to provide a framework for organizing what is known as well as for planning for additional research and study. To illustrate this organizing function, the literature on verbal analogies was reviewed with regard to what it has to say about sources of variation in the difficulty of this item type.

Analogies are perhaps the most widely studied of the various psychometric tasks. A number of investigators such as Sternberg (1977), Whitely (1976, 1977), and Pellegrino and Glaser (1980) have studied verbal or figural analogies extensively. The focus of much of this research, however, has been on identifying the cognitive processes involved in solving analogy items,

rather than on either the characteristics of the items which are associated with variability in item difficulty or the individual differences associated with success on these items. Sternberg (1977), for example, used analogy items which set cognitive processing demands well within the capabilities of the subjects he was using so that few if any solution errors occurred. In these studies, performance was measured in terms of response time.

Some research results and suggestions based on theoretical formulations can nonetheless be found in the literature regarding sources of item difficulty in analogies. The following review will be divided according to the two major sources of variation in item difficulty proposed in the theory: the features of the items and individual differences among examinees.

Sources of Difficulty in the Features of Analogy Items

One possible source of difficulty in analogy items is the nature of the semantic relationship between the two terms in the stem of the item. A number of classification schemes of verbal analogy items have been proposed (Whitely, 1977; Chaffin & Herrmann, in press a,b; Freedle & Gitomer, 1985). Although these classification schemes differ in terms of level of detail and inclusiveness, all include relationships such as synonyms, antonyms, part-whole relations, various functional relations (e.g., agent-action, action-object), and causal relations.

The difficulty of the analogy, however, is unlikely to be dependent only on the category of the correct relationship as given in the stem. Pellegrino and Glaser (1980) pointed out that the features associated with the set of alternatives can also affect the item difficulty. The relation between stem and options might best be described through the concept of similarity. Sternberg and McNamara (1985) discussed this concept in relation to difficulty

in synonym items. They indicated that the more dimensions along which the meaning of a stem and a response are similar, the more likely they are to be recognized as synonyms. The difficulty of the item may then be further affected by distractors. Where a word provided as a distractor has some dimensions of similarity with the stimulus word, but fewer than the key word, this distractor will be more difficult to eliminate from consideration as the correct response than words with few if any dimensions of similarity.

VanLehn and Brown (1981) described similarity in analogy items where the comparison was between the semantic relations rather than the individual words as was the case in synonym items. That is, the comparison was between the relation of the word pair provided as a stimulus in the item stem and the relation of the word pairs provided in the various options. The more dimensions along which the relationships within the stimulus and response parts of the analogy are similar, the more likely that a correct analogical relationship between the parts will be recognized. Likewise, a distractor which contains a relationship with some elements in common with the stimulus, but fewer common elements than the key, will be more difficult to eliminate.

Chaffin and Herrmann (in press b) elaborated this concept still further in defining heterogeneous and homogeneous items. They first empirically identified a small number of relationship "families" within each of which are some number of specific relationships. Homogeneous items are then those where all options are from the same relationship family, but only the key has the same specific relationship as the stem. In heterogeneous items, only the key belongs to the same family as the stem. Heterogeneous items can be further differentiated, however, according to whether or not the key also has the same specific relationship, although in both instances only knowledge of the

relationship family is required to select a correct response. Chaffin and Herrmann (in press a) demonstrated that these classifications, which consider the option set, were associated with the difficulty of the items, with homogeneous items generally more difficult than heterogeneous. The patterns of difficulty were not consistent over different relationship families, however, suggesting that other sources of difficulty were also operating in these items.

Among the other probable sources of difficulty in analogies are the salience or ambiguity of the relationship as provided in the item stem. Pellegrino and Glaser (1980) defined three experimentally determined variables for use in the study of this factor, which they refer to as the degree of semantic constraint. Subjects were asked to provide the relationship between two terms and the following were obtained for each pair: (a) the probability associated with the single most frequently generated response, (b) the probability that the generated response reflected use of an appropriate semantic relationship, and (c) the number of different responses generated. An item with a high level of constraint would be one with high probability that a single correct response will be generated. This would correspond to the type of analogy item where the correct response could be readily produced without reference to the options. A item with low constraints would be one where the options would need to be evaluated in order to identify the appropriate relationship in the stem from among several possibilities. Generally, a highly constrained item would be expected to be easier than one with little constraint.

Chaffin and Herrmann (in press a) similarly defined concepts of relation identification, expressibility, and ambiguity. In their terminology,

identification refers to the correct association of a given word pair with a particular relationship. The difficulty of identification of a relationship can be expected to vary over different word pairs. They found, for example, that errors and response latency increased as word pairs became increasingly unclear, though correct, examples of a stated relationship. Relationship expression refers to the way in which the relation is described in "common parlance." An item in which the stem relation can readily be expressed might be expected to be easier than one where the relation is less clearly specifiable. Ambiguity arises from two possible sources: (a) more than one meaning for one or both of the stimulus terms or (b) more than one relationship that can correctly be identified when the meaning of the stimulus terms is retained. An ambiguous relation would thus be expected to result in an item with low constraint.

Other sources of difficulty suggested by Pellegrino and Glaser (1980) include (a) the "conceptual richness and/or abstractness of the individual terms and relations" (p. 213), (b) the complexity of the relationship, and (c) the declarative knowledge base. This last would, of course, include level of vocabulary and other general knowledge usually assumed to be available to most examinees. An aspect of relation complexity, which might be expected to affect difficulty, was described by Chaffin and Herrmann (in press a) as part of their concept of relation creativity. Working with examples from the Graduate Record Examination (GRE), much more difficult forms of analogy items than have typically been studied in this literature, they described concatenations of relationships in which two terms are related through a linkage which is not stated but only implied in the item stem. That is, if one is to describe the relation of the stimulus word pair in terms of the

classification categories discussed above, more than one such relation is required through the insertion of an unstated link term. A simple example would be sailor:anchor, where the unstated term is ship.

Some additional possibilities for sources of difficulty arise from measurement practice and have not generally been discussed in the psychological literature. These include structural properties such as option order or the relative placement of the various terms of the analogy For example, test developers believe that for certain items, reversing the stem and key or the two terms in the stem would produce items with differing difficulty. (This belief may be well founded. Reversal of stem and key may easily result in changes in item properties such as those described in the paragraphs above.) The analogy format of A:B::__:__is typically used at Educational Testing Service, although an alternative form sometimes used elsewhere is A:B::C: . The latter format is often assumed to produce easier analogy items, other things being equal. Finally, another source of difficulty is item flaws. If an unintended relationship between the terms of the stem can be perceived by some examinees which can then be matched with one of the options intended to be a distractor, the item will be more difficult (i.e., fewer examinees will select the intended key) than would be the case without this flaw.

Sources of Difficulty in Individual Differences in Processing

A number of investigators have specifically studied individual differences in processing strategies in the solution of analogy items. Bisanz, Bisanz, and LeFevre (1984) defined strategies in terms of a particular combination of rules and attributes. Studying children at different age levels in order to observe the development of strategies, they found that

younger and less able children were often found to be applying non-analogical rules. Further they found individual differences in the extent to which the rules were applied to the task relevant attributes of the different terms of the analogy. The authors concluded that inadequate test directions may result in "failure to understand problem constraints and the consequent use of inadequate strategies" (p. 174).

Heller and Pellegrino (1978) used think-aloud procedures with a set of analogy items to identify eight different strategies to solution of analogies. Some of these strategies were much more likely than others to result in a correct solution so that the strategies chosen tended to be different for individuals with different analogical reasoning abilities. The same individual, however, might use different strategies with different items with varying degrees of success. The major difference between the strategies identified was whether the strategy concerned working forward or backward. In the first of these, the examinee worked forward from the stem, forming a hypothesis as to the appropriate relation and seeking a suitable example of that relation from among the options. In the backward strategies, the appropriate relationship was identified by considering the options and constraining the possible relations accordingly. Of course, the effectiveness of one of these strategies depended on the characteristics of the item being solved. More able examinees were more likely to adopt a strategy more appropriate to the item features.

Alderton, Goldman, and Pellegrino (1985) also found differences in the strategy used for the solution of analogy items by groups differing in ability. In addition to variables like those identified by Heller and Pellegrino (1978), they investigated "distractibility," that is, the tendency