ED 395 012 TM 025 026

AUTHOR	Olson, John F.; And Others
TITLE	Statistical Approaches to the Study of Item Difficulty.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-89-21
PUB DATE	Sep 89
NOTE	34p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Classification; Cluster Analysis; *Difficulty Level; Educational Research; *Measurement Techniques; Research Methodology; *Statistical Analysis; Teacher Certification; Test Construction; *Test Items
IDENTIFIERS	*Confirmatory Factor Analysis; Exploratory Data Analysis; *Full Information Factor Analysis; Local Independence (Tests); NTE Specialty Area Tests

ABSTRACT

Traditionally, item difficulty has been defined in terms of the performance of examinees. For test development purposes, a more useful concept would be some kind of intrinsic item difficulty, defined in terms of the item's content, context, or characteristics and the task demands set by the item. In this investigation, the measurement literature was surveyed for statistical approaches that might be applied to the study of item difficulty. Two broad methodological approaches were identified, exploratory and confirmatory approaches. Exploratory methods are those that attempt to categorize or cluster items that appear to measure similar abilities, that function in a similar manner in order to determine their common characteristics, and that differentiate them from other items not in the cluster. Confirmatory methods would be applied to test hypotheses developed from exploratory results or from psychological theory. The final section of the paper describes analyses using real test data that assessed the usefulness of two exploratory methods. Data from the NTE specialty area test for teacher certification in social studies for 1,748 examinees were used to evaluate full-information factor analysis and a measure of local item independence. The analyses indicate the usefulness of exploratory methods. (Contains 53 references.) (Author/SLD)

RESEARCH REPORT

STATISTICAL APPROACHES TO THE STUDY OF ITEM DIFFICULTY

John F. Olson
Janice Scheuneman
Angela Grima

Educational Testing Service
Princeton, New Jersey
Septemb 989

2

STATISTICAL APPROACHES TO THE STUDY OF ITEM DIFFICULTY

John F. Olson, Janice Scheuneman, Angela Grima

Educational Testing Service

August 1989

This work was supported by ETS through the Program Research Planning Council.

Abstract

Traditionally, item difficulty has been defined in terms of the performance of examinees. For test development purposes, a more useful concept would be some kind of intrinsic item difficulty, defined in terms of the item's content, context, or characteristics and the task demands set by the item. 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. An important step in developing the knowledge base required to acquire an understanding of those item properties that affect difficulty is appropriate analyses of existing test data. In this investigation, the measurement literature was surveyed for statistical approaches which might be fruitfully applied to the study of item difficulty. Two broad methodological approaches were identified: exploratory and confirmatory approaches. Exploratory methods were those that attempt to categorize or cluster items that appear to measure similar abilities and that function in a similar manner in order to determine their common characteristics as well as those that differentiate them from other items not in the cluster. Confirmatory methods would be applied to test hypotheses developed from exploratory results or from psychological theory. Described in the final section of the paper are the results of analyses using real test data that assessed the usefulness of two of the exploratory methods.

STATISTICAL APPROACHES TO THE STUDY OF ITEM DIFFICULTY

John F. Olson, Janice Scheuneman, Angela Grima

Educational Testing Service

Traditionally, item difficulty has been defined in terms of the performance of examinees. Classical theory has defined difficulty as the proportion of examinees responding correctly to the item or some transformation thereof. Item response theory (IRT), while freeing item statistics from the peculiarities of particular samples of examinees, still defines difficulty in terms of the probability of a correct response at a given level of examinee ability. Little attention has been paid to the intrinsic difficulty of an item, that is, to item difficulty defined in terms of the content or other properties of the item.

Intrinsic item difficulty would be defined in terms of the content, context, characteristics or properties of the item 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 (McGrail, Scheuneman, Steinhaus, & Swinton, 1988).

Appropriate analyses of existing test data are an important step in developing the knowledge base needed to acquire an understanding of those item properties which affect difficulty. Fundamentally, we expect that item difficulty functions primarily as a result of the material being tested, but, in fact, experienced test developers report themselves able to write easy items concerning difficult material and difficult items about easy material (McGrail et al., 1988). Scheuneman, Gerritz, and Embretson (1989) found that measures of prose complexity added significantly to the prediction of item difficulty beyond that provided by measures representing the knowledge requirements of the items. Ideally, then, analyses reveal not only the effects on difficulty of components of the knowledge, skill or ability domain a test is intended to measure, but also the effects created by item demand on other domains or irrelevant sources of difficulty introduced by properties of the surface structure of the items (Scheuneman & Steinhaus, 1987).

In this investigation, the measurement literature was surveyed for statistical approaches which might be fruitfully applied to the study of item difficulty. Various techniques are discussed which attempt to categorize items that appear to measure similar abilities and that function in a similar manner in order to determine the characteristics that the items in a cluster appear to have in common but that differentiate them from other items not in the cluster.

Two methodological approaches to the study of sources of variation in item difficulty might be distinguished as exploratory and confirmatory approaches. The first of these would approach the problem in a strictly empirical way. First, clusters of items would be identified which appear to function in the same way in contrast to other items on the same test, as reflected in examinee performance. Once such clusters were identified, they could be examined by subject matter and measurement experts in order to determine the properties of the items or the processes required to solve them that would distinguish one cluster of items from another. From such evaluations, hypotheses concerning possible sources of difficulty could be formed. These empirical hypotheses could then be combined with others from previous research and evaluated in a confirmatory mode. In general, confirmatory studies are designed to evaluate specific hypotheses concerning sources of item difficulty.

In this paper, various methodologies which might be appropriate for clustering items in an exploratory mode are reviewed in the first section and possible methodologies for use in confirming specific hypotheses or models of difficulty are discussed in the second section. In thε third section, the results of studies conducted by the authors to assess the usefulness of two of the exploratory analyses discussed in this paper are presented.

Exploratory Methods for Forming Item Clusters

A major way in which items will cluster is according to differences in the specific or component abilities measured by the item sets. Thus, one way of identifying these clusters is by using one of the methods designed to determine the underlying dimensionality of the test item data. Some of the approaches that have been used for this purpose and that are discussed in this section are:

- factor analysis
- cluster analysis
- order analysis
- investigation of item response patterns
- tests of local independence

Factor Analysis

Factor analysis is a technique commonly used to assess the dimensionality of data. This method assumes that the observed variables are linear combinations of some underlying factors or constructs and that the variables are measured at least at the interval level. A problem often exists in factor analysis applications of item data since items are usually scored at a dichotomous level of right or wrong. For instance, Carroll (1945, 1961, 1983) documented the problems inherent in the factor analysis of phi coefficients. He points out that such correlations depend not only on the strength of the relationship between the variables being correlated, but also upon their means. Mislevy (1986) warns against analyzing phi coefficients which may be dichotomized at different points. He notes that they may conform to factor models with different structures and possibly different numbers of factors. In their research, McDonald and Ahlawat (1974) tried to explain the existence of "difficulty factors." Carroll (1945, 1961, 1983) tried solving the problem of difficulty factors by using tetrachoric correlations. He notes that unless guessing is taken into consideration and adjustments are made, artifactual factors may still emerge. And even then, further adjustments may still be needed (Mislevy, 1986; Hulin, Drasgow, & Parsons, 1983).

Due to the problems that occur with the use of dichotomous data, other approaches have generally been preferred for the purposes of clustering items. Recent developments in the factor analysis of categorical variables have been made, however, that extend the classical factor analysis methods to dichotomous test items (for example, see Mislevy, 1986). Some of the factor analytic methods that have been used to overcome these problems include the factor analysis of item parcels (Cook, Dorans, Eignor, & Petersen, 1985), non-linear factor analysis (McDonald, 1983), and item response theory based factor analyses. The latter include a generalized least squares approach (Christoffersson, 1975), a marginal maximum likelihood full-information factor analysis approach (Bock & Aitken, 1981) and Muthen's related procedures (1978, 1984). These methods appear to be promising approaches for the assessment of item data dimensionality by using factor analysis techniques. For the analysis discussed later in this paper, the item factor analysis was investigated in detail.

Bock, Gibbons, and Muraki (1986) present a detailed paper on the derivation of full-information item factor analysis and discuss some of the technical problems of using it as well as describing several of their applications with simulated and real data. Based on their research, they found item factor analysis to be the most informative and sensitive method for the investigation of the dimensionality of item data. The Bock and Aitken item factor analysis method is based directly on item response theory and does not require calculations of the inter-item correlation coefficients. However, like all the other IRT approaches, this method makes the assumption that the underlying traits are multinormally distributed. Researchers (Mislevy & Bock, 1983; Tucker, 1983) have tried to develop procedures that circumvent the multinormal distributional assumptions on the latent traits.

The TESTFACT computer program, developed by Wilson, Wood, and Gibbons (1984), uses a marginal maximum likelihood method to estimate the difficulty and discrimination parameters for a multidimensional IRT model. It does not require linear relationships among the data. The method provides a stepwise factor analysis to examine each factor for statistical significance as it is added to the model. Kingston (1986) used this procedure to assess the dimensionality of the Graduate Management Admission Test (GMAT) Verbal and Quantitative measures. In his research, Kingston found it to be a useful analytical technique because of its direct nonlinear factor analytic approach and because it provided a statistical test for the determination of a multidimensional factor model.

Cluster Analysis

Another way to classify and categorize data is by using a clustering methodology. Milligan and Cooper (1987) identify and describe four major types of clustering methods: hierarchical methods, partitioning (nonhierarchical) algorithms, overlapping clustering procedures, and ordination techniques. Hierarchical clustering methods seem to be the most popular and widely used approach. This technique is based on an agglomerative hierarchical clustering procedure where each observation begins as a cluster by itself, then the closest two clusters are merged to form a new cluster, and this process is repeated until only one cluster is left. In their paper, Milligan and Cooper discuss some of the advantages and disadvantages of using the various clustering techniques and recommend that the type of method used be dependent on the kind of data to be analyzed, the selection of the variables to be used in the cluster analysis, and the characteristics of the population.

An example of a cluster analysis application to test item data was done by Oltman, Stricker, and Barrows (1988) in their research on the structure of the Test of English as a Foreign Language (TOEFL). They investigated how level of proficiency in one's foreign language and in the English language interrelated with the structure of the test. A multidimensional scaling approach was used that accounted for individual differences in language proficiency and how the differences related to the number of dimensions that could be determined in the set of items. Then, the stimulus coordinates from the scaling analysis, which represent the item's locations on the different dimensions that were identified, were cluster analyzed using a hierarchical method in order to determine how the items were grouped together in the space defined by the dimensions. They found that the easier items in each section of the test tended to define the clusters and that the more difficult items did not fit well into any of the dimensions identified in the test. Their results indicated the dimensionality of the TOEFL depends on the level of English proficiency of the examinees, with more salient dimensions found in the least proficient populations of test takers. They concluded that the easy and difficult items were different in their ability to measure overall language proficiency and specific language skills, with easy items better suited for diagnostic purposes such as measuring specific language skills, and difficult items better measures of general proficiency and, therefore, more useful for global screening purposes. The authors suggest it may be possible to alter the content specifications of the TOEFL by changing the difficulty of the items in the test. They also strongly recommended the methodological procedures used in their research, advocating an increased use of multidimensional scaling and cluster analyses in the study of test data.

Order Analysis

Krus and Bart (1974) presented a method for multidimensional scaling of dichotomous item data that was derived from ordering theory. This method is related to a multivariate extension of Guttman's scalogram analysis technique. Krus and Bart applied this method to item data response patterns from a hypothetical set of data used in a previous study. This approach is somewhat analogous to factor analysis but does not employ correlational procedures. The authors state that this method can be a very useful technique in that it can be used to scale any set of test items in a multidimensional manner and can also determine the number of dimensions in the data, using the rank ordering loading matrix as a multidimensionality indicator.

Krus (1977) used an order analytic approach to derive an inferential model for multidimensional analysis and scaling. He used the McNemar z statistic to evaluate the presence of any dominance relations in a collection of items. This approach utilized a deterministic order analytic and probabilistic model to generate order loadings for the items on each dimension. Krus (1978) followed this work with a further application of order analysis. This technique was developed as a method of multidimensional analysis and scaling based on the theory of Boolean algebra. In an examination of the Marital Adjustment Inventory, five dimensions were found. Krus then compared and contrasted the order analysis approach with the principal factors method of factor analysis. Moderate structural similarities were found between the two approaches. The difference between the two techniques is that order analysis is designed for the analysis of matrices of dominance coefficients and utilizes functions of the propositional calculus, whereas factor analysis focuses on the analysis of matrices of correlation coefficients.

Reynolds (1981) utilized a method called ERGO, which is based on the logic of ordering theory. This method extracts reliable item hierarchies of the Guttman type. It was applied to an investigation of the dimensionality of the Social Distance Questionnaire with multiple ethnic groups. It differs from factor analysis as a clustering technique in that it takes item difficulty into account. Reynolds found this method superior to factor analysis in that it obtained an hierarchical-developmental ordering of the items.

Wise (1983) investigated the use of proximity measures and compared the use of factor analysis and order analysis to assess the dimensionality of binary data. The data were of a known dimensionality. Wise compared the Krus and Bart (1974) method of order analysis with two order analytic approaches used by Reynolds (1981) and found the Krus and Bart method and Reynold's extraction index method to be poor methods of determining dimensionality for the datasets that Wise was analyzing. Reynold's other order analytic method (C_{t3}) was found to be useful with datasets consisting of orthogonal factors but not with oblique factors. In a more recent study, Wise and Tatsuoka (1986) demonstrated that using the proximity information to modify the order analysis procedures yielded results that were congruent with those from factor analysis.

Investigation of Item Response Patterns

Although ordering theory methods use item response patterns, they differ from the techniques in this section in that the methods described here identify dimensionality by highlighting persons or groups of persons rather than clusters of items.

There are two major sets of indices which are useful in determining the degree to which an individual's pattern of item responses is found to be unusual. One set of indices are based on item response theory. These include the "appropriateness" indices described by Levine and Rubin (1979) and later modified by Drasgow (1978, 1982). The chi-square test of person fit which is used in applications of the Rasch model (Wright, 1977) is also an IRT-based index.

The second set of indices, group-dependent indices, are based on the pattern of right and wrong answers. Among these are the "caution" index (Sato, 1975), a modified "caution" index (Harnisch & Linn, 1981), the "U" index (Van der Flier, 1977), the norm-conformity index (Tatsouka & Tatsouka, 1982), and agreement and disagreement indices (Kane & Brennan, 1980).

Harnisch (1983) utilized item response patterns to identify individuals with unusual response patterns on achievements tests. The approaches used in his research were conceptualized from Student-Problem (S-P) curve theory (Sato, 1975). The approaches used include the ability to overcome limitations of global summary scores, especially with tests consisting of interrelated subsets of achievement test items, and to identify distinct response patterns to assist in the analysis, interpretation, and reporting of achievement data. This type of approach can also aid in the determination of whether a collection of items or subjects form a heterogeneous group.

Mayberry and Ory (1985) used a related technique to cluster persons with related abilities or misconceptions of the subject matter based on their item response patterns. In this procedure, they plotted IRT ability estimates (based on a 2-parameter logistic model) against an "extended caution index" based on the S-P chart conceptions. This enabled them to identify students with similar strengths and weaknesses and hence to identify some of the component abilities measured by the test.

Tests of Local Independence

One of the underlying assumptions of IRT models is the assumption of unidimensionality. This assumption implies that the items measure one and only one area of knowledge or ability. If it is satisfied, then the assumption of local independence is also met. There are two forms of local independence, strong and weak. The former states that an examinee's responses to different items on a test are statistically independent at a given level of ability. That is, an examinee's performance on one item must not affect, in any way, his or her responses to any other items on the test. The probability of any pattern of item scores occurring for an examinee is thus equal to the product of the probability of occurrence of the scores on each item (Hambleton & Swaminathan, 1985). The weak form of local independence states that at a given ability level, an examinee's response to one item is uncorrelated with the response to any other item.

As already noted, an important assumption of IRT models is that responses to the items are locally independent. However, to the extent that the unidimensionality assumption is not met, some dependence among items may arise because they measure an unintended ability which varies for persons who are equivalent on the intended ability. As a result, measures of local independence which have been developed to test the fit of the data to this assumption may also be sensitive tests of multidimensionality.

Several researchers have investigated various techniques for testing the assumption of local independence (e.g., Kingston & Dorans, 1982; Kingston, Leary, & Wightman, 1985; Yen, 1984). In her research, Yen (1984) investigated the use of several measures of fit for the examination of the effects of local item independence toward utilization of the three parameter logistic model for equating. She analyzed both real and simulated data.

The first measure, Q_1, consists of a comparison between observed and predicted item characteristic curves. Although this statistic is only a goodness of fit measure, we do know that one of the factors that can affect the fit of the model is multidimensionality. Hence, if the item does not fit the model, one may then question the assumption of unidimensionality.

The second statistic, Q_2, is a generalization of Van den Wollenberg's (1982) fit measure for the Rasch model. Although this statistic is useful in determining where local independence exists, when violations occur, it does not reflect whether they are in a positive or negative direction. Therefore, in order to estimate the direction of the relationship between the items, a "signed Q_2" statistic was also derived and utilized.

The third statistic, Q_3, a revised version of the statistic used by Kingston and Dorans (1982), assesses the correlation of item scores with the ability trait estimates partialed out. Kingston and Dorans used the earlier version of Q_3 to test the weak form of the local independence condition for the feasibility of using IRT as a psychometric model for the Graduate Record Examinations (GRE) General Test. Although they were satisfied with the obtained results, as Yen (1984) points out, their statistic has one disadvantage; it is only capable of removing the linear relationship between item scores and traits when it is well known that a nonlinear logistic relationship probably exists. On the other hand, Yen's alternative measure, removes the nonlinear effects of the ability trait estimate from the item scores.

In Yen's research, the results show that Q_1 had low correlations with Q_2 and Q_3. In addition, the factors which cause misfit as measured by Q_1, do not appear to include multidimensionality. Previously, Yen (1981) had noted that Q_1 was not useful in determining when a two-parameter model was inappropriately applied to three-parameter data. Thus, she concluded that although it can be useful in identifying items that have unexpected characteristic curves, it cannot be relied upon as a complete fit measure. On the other hand, the results obtained for Q_2 and Q_3 were found useful for identifying subsets of items that were influenced by the same factors or that had similar content.

Another group of researchers (Kingston, Leary, & Wightman, 1986) conducted an exploratory study of the applicability of IRT methods to the GMAT in which they used a number of methods for assessing the reasonableness of the local item independence assumption and the fit of the IRT three-parameter logistic model to their data. One of the approaches used was a modified Q_1 statistic, a revised version of a measure evaluated by Yen (1981). Unlike the earlier version of Q_1, which used ten groups with approximately equal sample sizes, the revised statistic uses seventeen groups based on equal intervals along the ability metric. In using this statistic to assess the assumption of local independence, these researchers considered the probability of Type I error rather than the statistic itself. The results observed by using this technique were found to be consistent with those obtained from their other analyses.

Confirmatory Methods for Evaluating Hypotheses

Another means of investigating the effects of different ability dimensions on variation in item difficulty is to decide *a priori* what these dimensions might be and to evaluate whether items differing in their demand on these abilities in fact differ in their difficulty or discrimination. The first part of this section reviews several judgmental procedures for defining the ability dimensions in item sets. The second part of the section reviews a small number of studies which have used mathematical procedures specifically designed to evaluate hypotheses concerning sources of item difficulty.

Judgmental Methods

Macready (1983) discussed the use of generalizability theory to assess relations and groupings among items within domains in diagnostic testing. This method uses an ANOVA approach for the assessment of generalizability (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), and is based on conducting a logical analysis to determine the underlying skills necessary for adequate performance on the test. Macready's investigation examined item homogeneity in a domain-referenced test, the Arithmetic Test Generation Program, which deals with the multiplication of whole numbers. The author states that this logical approach used to define the domains in the content area being investigated provided a reasonable initial approximation to the desired groupings of the items, but that additional research was needed to further assess the capabilities of this method.

Kolen and Jarjoura (1984) described an approach to analyzing items which is appropriate for the heterogeneous nature of several achievement and professional certification tests. This approach, called item profile analysis, compares the profiles of observed and expected correlations of item scores with category (based on content) scores in order to determine the fit of an item to a content category. The concept of a profile of expected correlations is derived from the model of generalizability theory which provides the basis for this approach. As an illustration of the analysis technique, Kolen and Jarjoura used data from a professional certification program and attempted to link test development issues to generalizability theory. In conclusion, they recommend that item profile analysis should be used in addition to standard statistical procedures, especially with tests that are known to have a heterogeneous content.

Hartke (1978) investigated the use of latent partition analysis as a technique to test for a conceptually homogeneous item population. Hartke describes this method as a "logical judgmental process" whereby a group of knowledgeable individuals evaluate the item population and partition it based on the different skills or knowledge required by the examinees to respond correctly to the items. The technique was applied to an elementary algebra test. The author states that latent partition analysis determined a consensus of several sorters (evaluators) without limiting the nature or number of partitions identified by each sorter, and that the technique can be made to be an empirical methodology.

Predictions of Item Difficulty

In their investigation, Stenner, Smith, and Burdick (1983) first developed a theory of receptive vocabulary which hypothesized a number of specific relationships between item difficulty and some characteristics of the words used as stimuli for the items of the Peabody Picture Vocabulary Test. These hypotheses were then evaluated by determining whether indicators of the item characteristics did in fact predict item difficulty in this data set. Item difficulty was expressed in different metrics and standard multiple regression methods were used. This procedure was also adopted by Smith and Green (1985) in predicting difficulty of items from properties of the stimulus on a paper-folding test. Both studies showed that such predictions could be made.

A more elaborate statistical procedure was used by Embretson and Wetzel (1987) to evaluate a number of different models of prose complexity. This procedure consists of a comparison of the fit of the different models to a null model which assumes that all items have the same difficulty and a "perfect" model (in this case, the Rasch model) which contains a separate difficulty for each item. The improvement of the fit with a particular model of interest over the null model would be considered an indication that some sources of item difficulty are being accounted for by the variables specified in the model. The percent of variance in difficulty accounted for by these variables can be estimated from the results obtained.

Reckase (1985) described a multidimensional measure of difficulty based on a generalization of item response theory concepts and applied it in a study of the ACT Assessment Mathematics Usage Test. This measure provides a way of determining the difficulty of an item that can give useful information when test items measure more than one ability or dimension. Since this approach assumes that the item is of a known dimensionality, other techniques need to be applied first to assess the dimensionality of the test. The indices may then be used to observe the effects of the multidimensionality on observed item discrimination values, such as biserial correlations, based on a total score in which the different dimensions are combined and confounded.

Empirical Studies

In order to evaluate two of the procedures discussed in the literature review, a recent form of the NTE Specialty Area test used for teacher certification in Social Studies was examined. This test was of interest because of the possible heterogeneity of its content. The test consists of 150 five-option multiple-choice items measuring knowledge primarily in the Social Studies domain, of which, 149 items were scored. It was administered to 1748 examinees in April 1985.

Full-Information Factor Analysis

An investigation of the test data was conducted using full information factor analysis. The data were analyzed to assess its dimensionality prior to the estimation of parameters for a three-parameter logistic model. The TESTFACT computer program was run using a non-linear, maximum likelihood approach that is appropriate with dichotomous data such as the right/wrong scoring of test items. A TESTFACT factor analysis proceeds via a three-parameter multidimensional normal ogive IRT model. Based on the results provided by the computer output, one major factor was found that accounted for approximately 14.7% of the total variance in the test. The next largest factor only accounted for approximately 3.3% of the total variance, and a third factor accounted for 1.6% of the total variance.

Next, an oblique rotation of the factor loadings was made to assist in the interpretation of the factors. An examination of the content of the factors was done by determining which items loaded on each factor and then inspecting the items within each factor to see what they had in common. Based on an inspection of the content of the items within each factor, the primary factor was found to consist mainly of items that were measuring concepts related to the topics of American history and government. The second factor appeared to consist mainly of items that covered content areas related to basic concepts in sociology and social studies, and also the knowledge of basic teaching principles ("Professional Information"). The third factor was found to contain items related to world history, data reading, and a variety of miscellaneous topics related to social studies (e.g., geography, economics, political science).

The three factors were all correlated with each other, as can be seen in the following table:

PROMAX FACTOR CORRELATIONS

	1	2	3
1	1.000		
2	0.456	1.000	
3	0.605	0.608	1.000

An analysis of the latent roots of the tetrachoric correlation matrix was then conducted using a scree test which examines the latent roots used in determining the number of significant factors. The relative strengths of the factors indicate the test's dimensionality. The values for the three largest latent roots of the correlation matrix that were examined by the scree test are as follows:

LARGEST LATENT ROOTS OF THE CORRELATION MATRIX

FACTOR		
1	2	3
35.82	3.26	2.32

(all other latent roots had values less than 2.00)

As can be seen in this table, the first root is about 11 times larger than the second root, and the second root is less than twice as large as the third and not much larger than the remaining roots. This comparison of the magnitudes of the three largest latent roots shows that the first factor was by far the largest and most important factor. Thus, the scree test resuits suggest that the test may be reasonably one-dimensional for the purposes for which IRT models are typically applied.

Although a factor analytic model containing three factors was examined in detail, it must be noted that the amount of variance accounted for by this model was less than 20 percent of the total variance in the test, and the largest factor only accounted for about 15 percent of the total variance. This leaves a large proportion of the amount of information that is measured in the overall test unaccounted for. An appropriate interpretation is that this test contains more variance specific to the individual items than can be attributed to the factor structure; therefore, the test appears to be a rather heterogeneous measure. These results suggest that some other abilities are being measured which were not statistically derived by this factor analysis method.

Note that there are some limitations with using a full-information factor analysis approach. The TESTFACT program can be rather expensive to run, especially when testing for as many as four or five factors. The program can require a substantial amount of processing (CPU) time in order to complete its iterative computations. For this reason, the maximum number of factors that were tested for statistical significance was held to three for this study. Although this approach does not appear to be very sensitive for determining possible variations in item content or type, it may be useful for an initial exploratory analysis of the overall structure of the data prior to using other procedures.

Local Item Independence

Since the data were already available as output from the IRT calibrations for the Social Studies test, a measure of local independence was also evaluated. The modified Q_1 statistics suggested by Kingston, Leary, and

Wightman (1985) were calculated and the probabilities of the Type I errors for the Q_1 values were tabulated. The following table presents the distribution of the probability of the Q_1 statistics, $P(Q_1)$, grouped into five classification ranges: .00-.05, .06-.25, .26-.50, .51-.75, .76-1.00. The statistics are shown within each content category and over the total test. Low values for $P(Q_1)$ indicate a poor fit of the test data to the three-parameter logistic model. Proportions of the items falling in each category are indicated in parentheses.

DISTRIBUTION OF PROBABILITIES ASSOCIATED WITH Q_1

	$P(Q_1)$					
	.00-.05	.06-.25	.26-.50	.51-.75	.76-1.00	TOTAL
Professional Education	0(.00)	7(.35)	3(.15)	4(.20)	6(.30)	20
Political Science & Economics	1(.03)	10(.28)	11(.31)	7(.19)	7(.19)	36
Sociology, Anthropology, Psychology & Geography	2(.04)	12(.27)	11(.24)	10(.22)	10(.22)	45
History: American & World	2(.04)	7(.15)	12(.25)	12(.25)	15(.31)	48
Total Test (All Categories)	5(.03)	36(.24)	37(.25)	33(.22)	38(.26)	149

The values in the table do not show any violations of local item independence for any of the categories or for the test as a whole. In comparison to the results of the Q_1 analysis found by Kingston et al. (1985), the present analysis found a much better fit of the items to the model being investigated. The proportions of items falling within the various ranges of $P(Q_1)$ approximate the expected chi-square distribution with only 3 percent of the items found to have probabilities less than .05, whereas 12 percent of the GMAT items fell in the same low category. Therefore, based on the analysis of the modified Q_1 statistic, the test data appear to fit the three-parameter logistic model.

These results confirm the previous conclusion that the Social Studies test is sufficiently unidimensional for the use of IRT models, but do not reflect any of the possible variations in the abilities measured suggested by the low percent of variance accounted for by the factors emerging from the TESTFACT analysis. Q_1 does not appear to be a useful statistic for item clustering or as an aide to assist in the study of item difficulty. However, it was found to be a useful measure in identifying individual items that had unusual characteristic curves and, thus, failed to fit the three-parameter logistic model.

Summary and Conclusions

In this review, a number of statistical procedures were considered for their potential in illuminating the various facets of item difficulty. Procedures were roughly divided into those which are primarily exploratory and those which are confirmatory. The exploratory methods are largely those which explore the dimensionality of a test. These included methods, such as factor analysis of item data and tests of local independence, which have been developed in order to evaluate the unidimensionality assumptions required by many IRT models. Other exploratory methods were based on analyses of observed item response patterns rather than on application of mathematical response probability models as in IRT. The response-pattern methods include variations of ordering theory approaches, alone or in combination with factor analysis or other procedures. Confirmatory methods included both judgemental methods, some based on generalizability theory, and statistical procedures by which a *priori* hypotheses concerning the dimensionality of the data or sources of item difficulty or discrimination could be evaluated.

Data from the NTE Social Studies examination were then used to evaluate item level factor analysis (TESTFACT) and an index of local independence (Q_1). The social studies test was an interesting example because of the variety of academic disciplines touched on by the exam. Unfortunately, the results were disappointing. Neither of these procedures appeared sensitive to the variations in item content or other properties of interest. The TESTFACT program or similar analysis procedures may be useful, however, in forming initial item groupings which might then be explored further with other, possibly more sensitive, procedures. Item pattern methods, for example, are most suitable for relatively small item sets. An alternative conclusion is that atheoretic, exploratory approaches are not going to be useful for this purpose. Logical analyses may be required in order to develop specific testable hypotheses, which can then be evaluated using confirmatory methods.

The judgemental methods may be of particular interest in helping to develop testable hypotheses. Although these procedures are confirmatory, they may offer means of helping to articulate and evaluate the working knowledge of experienced test developers. Much of what test development experts "know" is almost intuitive and some of it may be wrong (McGrail et al, 1988). Nonetheless, this is a resource for the exploration of item difficulty which might be tapped using these procedures.

Once specific hypotheses for sources of item difficulty are formed, the statistical confirmatory methods then become appropriate. Embretson and Wetzel's (1987) sequential modeling procedure seems particularly promising. For investigations into variation in item discrimination, Reckase's (1985) multidimensional IRT approach may prove useful. As our knowledge grows, these procedures will also be applied and evaluated. In the long run, the statistical confirmatory approaches are likely to be the strongest tools in our quest to understand intrinsic item difficulty.

References

Bock, R. D. & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1986). Full-information factor analysis. Chicago: Methodology Research Center/NORC.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 10, 1-19.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. Psychometrika, 26, 347-372.

Carroll, J. B. (1983). The difficulty of a test and its factor composition revisited. In S. Messick and H. Wainer (Eds.), Principals of modern psychological measurement: A fetschrift for Frederic M. Lord. Hillsdale, NH: Erlbaum.

Christoffersson, A. (1975). Factor analysis of dichomotized variables. Psychometrika, 40, 5-32.

Cook, L., Dorans, N., Eignor, D., & Petersen, N. (1985). An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating (RR 85-30). Princeton, NJ: Educational Testing Service.

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

Drasgow, F. (1978). Statistical indices of the appropriateness of aptitude test scores. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. Applied Psychological Measurement, 6, 297-308.

Embretson, S. E., & Wetzel, C.D. (1987) Component latent trait models for paragraph comprehension tests. Applied Psychological Measurement, 11, 175-193.

Hambleton, R. K. & Swaminathan, H. (1985). Item Response Theory. Boston: Kluwer Nijhoff Publishing.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. Journal of Educational Measurement, 20, 191-206.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18, 133-146.

Hartke, A. R. (1978). The use of latent partition analysis to identify homogeneity of an item population. Journal of Educational Measurement, 15, 43-47.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones - Irwin.

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. Applied Psychological Measurement, 4, 1980, 105-126.

Kingston, N. (1986). Assessing the dimensionality of the GMAT verbal and quantitative measures using full information factor analysis. (RR 86-13). Princeton, NJ: Educational Testing Service.

Kingston, N. M., & Dorans, N. J. (1982). The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test (RR 82-12). Princeton, NJ: Educational Testing Service.

Kingston, N., Leary, L., & Wightman, L. (1985). An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test (RR 85-34). Princeton, NJ: Educational Testing Service.

Kolen, M. J., & Jarjoura, D. (1984). Item profile analysis for test developed according to a table of specifications. Applied Psychological Measurement, 8, 321-331.

Krus, D. J. (1977). Order analysis: An inferential model of dimensional analysis and scaling. Educational and Psychological Measurement, 37, 587-601.

Krus, D. J. (1978). Logical basis of dimensionality. Applied Psychological Measurement, 2, 323-331.

Krus, D. J., & Bart, W. M. (1974). An ordering theoretic method of multidimensional scaling of items. Educational and Psychological Measurement, 34, 525-535.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. Journal of Educational Statistics, 4, 269-290.

Macready, G. B. (1983). The use of generalizability theory for assessing relations among items within domains in diagnostic testing. Applied Psychological Measurement, 7, 149-157.

Mayberry, P. W., & Ory, J. C. (April 1985). Application of CAT misconception analysis in mathematics placement and proficiency decisions. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

McDonald, R. P. (1983). Exploratory and confirmatory nonlinear common factor analysis. In H. Wainer & S. Messick (Eds.) Principals of modern psychological measurement. Hillsdale, NJ: Erlbaum.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. British Journal of Mathematical and Statistical Psychology, 27, 82-99.

McGrail, E. C., Scheuneman, J. D., Steinhaus, K., & Swinton, S. (1988, November). A survey of experienced test developers concerning factors that influence item difficulty and discrimination (Statistical Report SR-88-154). Princeton, NJ: Educational Testing Service.

Milligan, G. W. & Cooper, M.C. (1987) Methodology review: Clustering methods. Applied Psychological Measurement, 11, 329-354.

Mislevy, R. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.

Mislevy, R., & Bock, R. D. (1983). Bilog: Item analysis and test scoring for binary logistic models. Chicago, IL: International Educational Services.

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 115-132.

Oltman, P.K., Stricker, L.J., & Barrows, T. (1988). Native language, English proficiency, and the structure of the Test of English as a Foreign Language. TOEFL Research Reports, No. 27, (RR-88-26). Princeton, NJ: Educational Testing Service.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. Paper presented at the American Educational Research Association annual meeting, Chicago.

Reynolds, T. J. (1981). ERGO: A new approach to multidimensional item analysis. Educational and Psychological Measurement, 41, 643-658.

Sato, T. (1975). The construction and interpretation of S-P tables. Tokyo: Meiji Tosho.

Scheuneman, J. D., Gerritz, K., & Embretson, S. E. (1989, March). Effects of prose complexity on achievement test item difficulty. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Scheuneman, J. D. & Steinhaus, K. (1987, December). A theoretical framework for the study of item difficulty and discrimination (RR-87-44). Princeton, NJ: Educational Testing Service.

Smith, R. M., & Green, K. E. (April 1985). Components of difficulty in paper-folding tests. Paper presented at the meeting of the American Educational Research Association, Chicago.

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. Journal of Educational Measurement, 20, 305-316.

Tatsouka, K. K., & Tatsouka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 7, 215-231.

Tucker, L. R. (1983). Searching for structure in binary data. In S. Messick and H. Wainer (Eds.), Principals of modern psychological measurement: A fetschrift for Frederic M. Lord. Hillsdale, NJ: Erlbaum.

Van der Flier, H. (1977). Environmental factors and deviant response patterns. In Y. H. Poortinga (Ed.), Basic problems in cross-cultural psychology. Amsterdam: Swets and Seitlinger, B. V.

Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. Psychometrika, 47, 123-140.

Wilson, D., Wood, R. L., & Gibbons, R. (1984). TESTFACT: Test scoring and item factor analysis, User's Guide, Version 2.2. Chicago: Scientific Software.

Wise, S. L. (1983). Comparison of order analysis and factor analysis in assessing the dimensionality of binary data. Applied Psychological Measurement, 7, 311-321.

Wise, S. L., & Tatsuoka, M. M. (1986). Assessing the dimensionality of dichotomous data using modified order analysis. Educational and Psychological Measurement, 46, 295-301.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. Applied Psychological Measurement, 8, 125-145.

OLSON:\PRPC\RESEARCH.REP
8/7/89