DOCUMENT RESUME

ED 270 489 TM 860 372

AUTHOR	Hambleton, Ronald K.; Rogers, H. Jane
TITLE	Promising Directions for Assessing Item Response Model Fit to Test Data.
SPONS AGENCY	Air Force Human Resources Lab., Brooks AFB, Texas.
PUB DATE	Apr 86
GRANT	F33615-84-C-0058
NOTE	58p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Data Analysis; Educational Research; *Factor Analysis; *Goodness of Fit; *Latent Trait Theory; *Mathematical Models; *Research Methodology; Research Problems; Statistical Significance; Statistical Studies; Test Items
IDENTIFIERS	Ability Parameters; *Item Invariance; Item Parameters; *Unidimensionality (Tests)

ABSTRACT

The general goal of this paper is to help researchers conduct appropriately designed goodness of fit studies for item response model applications. The specific purposes are to describe: (1) an up-to-date set of promising and useful methods for addressing a variety of goodness of fit questions; and (2) current research studies to advance this set of methods. Promising goodness of fit methods are organized around three main categories: (1) checks on the extent to which test data fit model assumptions; (2) investigations of item and ability invariance; and (3) model-test data fit studies. Two current investigations of the researchers were also reviewed: the use of non-linear factor analysis to address the assumption of unidimensionality, and new methodologies for addressing questions of item invariance. The recommended strategy for assessing model-data fit is to accumulate a considerable amount of evidence that can be used to aid in the determination of the appropriateness of a particular use of an item response model. Since a researcher can not prove that a test measures a construct, it was concluded that the more evidence accumulated, the more informed the final decision will be about the use of an item response model. References, tables and figures are appended. (Author/PN)

Promising Directions for Assessing Item Response Model Fit to Test Data

Ronald K. Hambleton and H. Jane Rogers
University of Massachusetts at Amherst

Abstract

The overall goal of this paper is to help researchers conduct appropriately designed goodness-of-fit studies for item response model applications. The specific purposes are to describe (1) an up-to-date set of promising and useful methods for addressing a variety of goodness-of-fit questions, and (2) some of our current research studies to advance the set of methods.

Promising goodness-of-fit methods are organized around three main categories: (1) checks on the extent to which test data fit model assumptions, (2) investigations of item and ability invariance, and (3) model-test data fit studies. Two current investigations of the researchers were also reviewed: (1) the use of non-linear factor analysis to address the assumption of unidimensionality, and (2) new methodologies for addressing questions of item invariance.

4/9/86

Promising Directions For Assessing Item Response Model Fit to Test Data[1,2]

Ronald K. Hambleton and H. Jane Rogers
University of Massachusetts, Amherst

Introduction

Item response theory (IRT) is presently being used by most of the large test publishers, the Armed Services, many state departments of education, large school districts, and a variety of industrial and professional organizations. IRT is being used to construct achievement and aptitude tests, to study differential item performance, to equate test scores, and to provide the measurement theory for computerized adaptive testing. The many applications have been sufficiently successful that discussions of IRT have definitely shifted in recent years from considerations of model advantages and disadvantages compared to classical test models to considerations of such topics as IRT model selection, parameter estimation methods, approaches for assessing model fit, and the development of specific guidelines for particular IRT applications. In these areas, the issues and technology associated with item response theory are not fully developed and some

[1] This research was supported by a contract from the Air Force Human Resources Laboratory (F33615-84-C-0058). The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of the Air Force position, policy, or decision, unless so designated by other official documentation.

[2] A paper presented at a symposium entitled, "Current Issues and Developments in Item Response Theory" at the annual meeting of AERA, San Francisco, 1986.

controversies still remain (see, for example, Goldstein, 1980; Traub & Wolfe, 1981). This paper examines one of these issues, the assessment of model fit.

While the potential of item response theory for solving many testing and measurement problems now appears to be substantial, the success of any application is not assured simply by processing test results through one of the available item response model computer programs (e.g., BICAL or LOGIST). A poorly fitting model will not yield invariant item parameter estimates, or statistics that accurately describe the items.

Neither can it be assumed that because so many datasets have been fit by item response models in the past that the fit to new datasets is assured. IRT applications in the measurement literature and especially the applications described in the large set of conference papers over the last ten years have often failed to address adequately the goodness-of-fit issue and so the extent of model-data fit is unknown (Divgi, 1981).

The advantages derived from the application of an item response model cannot be achieved when the fit between the model and the test data set of interest is less than adequate. Typical goodness-of-fit evidence presently involves statistical tests, but these tests cannot be used as the sole determiner of model-data fit because of their dependence on examinee sample sizes. When sample sizes are large nearly all departures from a model (even those where the practical significance of the deviation is minimal) will lead to rejection of the null hypothesis of model-data fit. With nall sample sizes even big differences may not be detected via statistical tests because of the low level of statistical power. In addition, the sampling distributions of some of the popular statistics are not as specified, and so errors will be made when these statistics are applied. This point is furthnr addressed in a later section.

With the goal of helping more researchers conduct appropriately designed goodness-of-fit investigations, the principal purposes of the present paper are to describe (1) an up-to-date set of promising and useful methods for addressing goodness-of-fit questions about item response models, and (2) some of our current research studies to advance the set of methods. Specifically, purpose one will be an update of some earlier work by Hambleton and Murray (1983) and Hambleton and Swaminathan (1985). Our review will include statistical tests, though other approaches seem more useful at present. Purpose two will be accomplished by presenting some of our recent work concerning statistical tests of fit, and investigations of unidimensionality and item parameter invariance.

Promising Methods for Addressing Goodness-of-Fit Questions

Overview

After a review of the model fit literature, Hambleton and Swaminathan (1985) suggested that the determination of how well an item response model fits a set of test data be addressed in three ways:

a. Determine the extent to which the test data satisfy the assumptions of the test model of interest.

b. Determine the extent to which expected advantages derived from the use of the item response model (i.e., invariant item and ability estimates) are obtained.

c. Determine the closeness of fit between predictions assuming the validity of the model and observable outcomes (for example, residuals and test score distributions) tilizing model parameter estimates and the test data.

For each of these approaches Hambleton and Swaminathan (1985) prepared a list of promising methods for collecting appropriate information. Figure 1 is an update of their earlier work. Basically, their orientation was to recommend that researchers avoid making important decisions based upon a narrow range of information. They recommended, instead, that researchers collect a wide range of data to influence the final judgment about model data fit. Checks on model assumptions and invariance properties, along with practical investigations of misfit and the consequences of misfit for the intended applications are all integral parts of the goodness-of-fit investigation.

Checking Model Assumptions

Item response models are based on strong assumptions which will not be completely met by any set of test data (Lord & Novick, 1968; Lord, 1980). There is some evidence that the models are robust to moderate departures, but the extent of robustness of the models has not been fully established (Hambleton et al, 1978). Given doubts about the robustness of the models, a researcher might be tempted to simply fit the most general model since it will be based on the least restrictive assumptions. Unfortunately, the most general models are multidimensional (i.e., assume that more than one latent variable is required to account for examinee test performance); these models are complex and do not appear ready for wide-scale use. Moreover, interpretation of results is complex and may not be what the researcher is looking for. Certainly, a multidimensional representation of ability is not common output from a test administration. Alternatively, it has been suggested that the three-parameter logistic model, the most general of the unidimensional models in common use, be adopted for important applications. The three-parameter model should result in better fit than either the one- or two-parameter models. There are three problems with this course of action: (1) more computer time is require⁴ to conduct the analyses, (2) somewhat larger samples of examinees and items are required to obtain satisfactory item and ability estimates and (3) the additional item parameters (item discrimination and pseudo-chance levels) complicate the use of the model for practitioners (see also Baker, 1986).

Model selection can be aided by an investigation of four principal assumptions of several of the popular item response models: unidimensionality, equal discrimination indices, minimal guessing, and non-speeded test administrations. Approaches for studying these assumptions are summarized in Figure 1. These approaches are considered in more detail by Hambleton and Murray (1983) and Hambleton and Swaminathan (1985).

There are many definitions of unidimensionality and this is one reason for confusion about assessing its presence. McDonald (1980, 1982) and Hattie (1981) arrived at the conclusion that the principle of local independence should be the basis for a proper definition for the assumption of unidimensionality. McDonald defined a set of test items as unidimensional if, for examinees with the same ability, the covariation between items in the set is zero. Since the relationship between items is typically non-linear, he recommended the use of non-linear factor analysis to study item interrelationships. Also, after fitting a single non-linear factor model to the item set, he recommended that residual covariances be calculated and used to assess the plausibility of the unidimensionality assumption. McDonald argued that the dimensionality of a set of test items should be determined by the number of factors or abilities needed for describing examinees in order to satisfy the principle of local independence.

Checking Model Features

Three measurement advantages are obtained when an item response model fits a set of test data:

1. Ability estimates are on the same ability scale and can be compared even though examinees may have taken different sets of test items from the pool of test items measuring the ability of interest.

2. The item statistics do not depend upon the sample of examinees from the population for whom the test is intended.
3. An indication of the precision of ability estimates at each point on the ability scale is obtained.

Item response models are often chosen as the mode of analysis in order to obtain these three advantages. However, whether or not these features are obtained depends on several factors -- model-data fit, test length, and precision of the item parameter estimates, among others. Through some straightforward methods, these features can be studied and their presence in a given situation determined.

The presence of the first advantage can be addressed, for example, by administering to examinees two or more samples of test items which vary widely in difficulty (Wright, 1968). It is most common to conduct this type of study by administering both sets of test items to examinees within the same test. Then, scores are obtained based on, say, the easier and harder halves of the test. Pairs of ability estimates obtained from the two halves of the test for each examinee are plotted on a graph. The bivariate plot of ability estimates should be linear, because expected ability scores for examinees do not depend upon the choice of test items when the item response model under investigation fits the test data. Some scatter of points about a best fitting line, however, is to be expected because of measurement error. When a linear relationship is not obtained, one or more of the underlying assumptions of the item response model under investigation are being violated by the test data set.

One weakness of the approaches described above is that there is no baseline data available for interpreting the plots. How is one to know whether the amount of scatter is acceptable, assuming model-data fit? An alternative is to perform statistical tests to study the differences between, say, b values obtained in two different subgroups. However, as long as there is at least a small difference in the true parameter values in the samples, statistically significant differences will be obtained when sample sizes are large. Thus, statistically significant differences may be observed even when the practical significance of these differences is inconsequential.

In a promising line of research for checking the invariance property, Hambleton and Rogers (1986) and Hambleton, Rogers, and Arrasmith (1986) generated a plot of b-values for randomly equivalent groups and then compared the plot to the plot of b-values obtained between two subgroups who might be expected to respond differently to some of the items (e.g., males versus females, Blacks versus Whites). The first plot serves as a baseline for interpreting the second plot. If the plots are similar, the groups are randomly equivalent and so the subgroups (e.g., male and female) are operating no differently from the randomly equivalent groups. Sex can then be ruled out as a factor in item performance. If the plots are different, attention shifts to identifying those test items which showed consistent differences in the subgroups of interest. The methodology for the "plot method" will be expanded on later in this paper.

Checking Additional Model Predictions

Several approaches for checking model fit are listed in Figure 1. One of the most promising approaches for addressing model-data fit involves the use of residual analyses. An item response model is chosen; item and ability parameters are estimated; and predictions of the performance of various ability groups on the items in the test are made, assuming the validity of the chosen model. Comparisons of the predicted results with the actual results are made (see, for example, Hambleton & Swaminathan, 1985; Kingston & Dorans, 1985).

By comparing the average item performance levels of various ability groups to the performance levels predicted by an estimated item characteristic curve, a measure of the fit between the estimated item characteristic curve and the observed data can be obtained. This process, of course, can be and is repeated for each item in a test. Most of the statistical tests of model fit that have been proposed use this approach.

In addition to this approach, it is reasonable and desirable to generate testable hypotheses concerning model-data fit. Hypotheses might be generated because they seem interesting (e.g., Are item calibrations the same for examinees receiving substantially different types of instruction?) or because questions may have arisen concerning the validity of the chosen item response model and the testing procedure (e.g., What effect does the context in which an item is pilot-tested have on the associated item parameter estimates?). On this latter point, see, for example, Yen (1980). Kingston and Dorans addressed the question of item context on item statistics and implications for the use of the statistics.

Researchers should also be encouraged to consider when possible the consequences of misfit upon their results. Hambleton and Cook (1983), for example, considered the effect of using the wrong model to obtain ability estimates. They also looked at the role of test length and sample size on item parameter estimates and, in turn, the effect of these factors on the precision of test information functions. Other researchers have studied the problem of other types of errors on equating and adaptive testing (e.g., Kingston & Dorans, 1984).

Summary

Our literature review revealed a substantial number of methods for conducting goodness-of-fit studies, but there appears to be too much emphasis on statistical tests for determining model-data fit. As an alternative, the use of researcher judgment in interpreting statistical tests of fit (rather than through the use of critical values) and other model-data comparisons for two or more models fitted to the same set of test data seems more desirable. Perhaps the statistical approach can be replaced by the use of graphical methods, replications, cross-validation techniques, study of residuals, baseline results to aid in interpretations, the study of practical consequences of misfit, and so on. In the last section of the paper, a set of recommended steps is offered.

With respect to testing model assumptions, unidimensionality is clearly the most important assumption to satisfy. Many tests of unidimensionality are available but those which do not use correlations

(Bejar, 1980) and/or incorporate the analysis of residuals (McDonald, 1980) seem most useful. In category two (checking model features), there is a definite shortage of ideas and techniques. Presently, plots of item parameter estimates obtained in two groups are compared without the aid of any "baseline plots", or statistical tests are used to compare the two sets of item parameter estimates. Such tests are less than ideal for the reasons offered earlier. Several new methods seem possible and one or two will be introduced in the planning sections. In the third category (checking model predictions), a number of very promising approaches have been described in the literature but they have received little or no attention from researchers (exceptions include the work of Kingston & Dorans, 1984; 1985; and several simulation studies, for example, Ansley & Forsyth, 1983). An outstanding example of a model data fit study (focusing on category 3) was recently completed by Hills, et al., (1985). Perhaps the problem is due to a shortage of computer programs to carry out necessary analyses or to an over reliance on statistical tests. In any case the problem is likely to be overcome in the near future.

Statistical Tests of Model Fit

Statistical tests of model fit are typically chi-square tests entailing comparison of observed results with those expected assuming model validity. Most of the chi-square tests that have been proposed were developed for the one-parameter model, although recent work has produced extensions to the two- and three-parameter models (see, for example, Yen, 1981).

Three main approaches to the construction of chi-square tests of goodness-of-fit may be identified. The first of these and the most commonly used is based on a standardized difference of observed and expected results; the second uses contingency table data; and the third employs a likelihood ratio. These different approaches to some extent reflect different parameter estimation procedures; likelihood ratio tests are only possible when conditional maximum likelihood estimation is performed.

Chi-square tests of fit based on the standardized residual are of the form

$$y = \sum_{j=1}^{k} \sum_{i=1}^{m} \frac{((f_{ij}) - E(f_{ij}))^2}{Var(f_{ij})}$$

where f_{ij} is the frequency of correct responses to item j among persons with score i;

$E(f_{ij})$ is the corresponding expected frequency, equal to $n_i p_{ij}$, where n_i is the number of persons with score i and p_{ij} is the probability of success on item j for persons with score i, calculated from the model parameter estimates.

m is the number of score groups (usually $k-1$) and k is the number of items.

Since f_{ij} has a binomial distribution with parameter p_{ij}, the normal approximation to the binomial yields a residual which has an approximately unit normal distribution; squaring and summing over items and score groups results in a chi-square on (m-1)(k-1) degrees of freedom.

The chi-square statistic derived in this manner has been criticized on several grounds. When any of the $E(f_{ij})$ terms has a value less than one, the claim of a chi-square distribution is of dubious validity, since the standardized residual of observed and expected will not be normally distributed. When sample sizes are small or do not contain a sufficient range of ability, this problem becomes severe. On the other hand, when sample sizes are large, the statistic gains sufficient power to detect trivial deviation from the model. Hence, both small and large sample sizes can adversely affect the behavior of the statistic. Nevertheless, variations on this statistic have continued to be used in the absence of better tests.

Wright and Panchapakesan (1969) first proposed the statistic described above as an overall test of the fit of the one-parameter model. Summing over only items or persons results in statistics which have been used to assess the fit of individual persons and items. Wright and Stone (1979) suggested the use of a variation on this form, called the mean square residual, as a measure of the fit of persons and items. Unlike the Wright-Panchapakesan statistic, the mean square residual is computed using individual responses rather than grouped responses.

The mean square residual is calculated by means of the formula

$$v = \sum \frac{(x_{ij} - p_{ij})^2}{p_{ij}(1-p_{ij})}$$

where x_{ij} is the response of person i to item j

(=1) if correct, =0 if incorrect)

and p_{ij} is the probability of success on item j for person i,

calculated from the model parameter estimates.

The statistic may be summed over items for person fit or over persons for item fit. It is unlikely that this procedure will result in a chi-square statistic, however, given that each residual is based on a single observation and cannot be normally distributed. Wright, Mead and Bell (1979) point out that the mean square residual is very sensitive to unexpected responses, such as correctly guessed answers for a person of low ability, and modify the statistic to "fr~tify" it against such responses. The statistic they produce is called the total-t, and is included in the BICAL program as a measure of the fit of persons and items. The total-t differs from the mean square residual in that the numerator and denominator are summed separateiy. The statistic produced is given by the formula

$$v = \frac{\sum (x_{ij} - p_{ij})^2}{\sum p_{ij}(1-p_{ij})}$$

where summation is over either person or items.

Although the statistic appears to be a ratio of variance estimates, its distribution is not clear. Wright, Mead, and Bell (1979) apply a cube-root transformation to produce a statistic which is said to be approximately normally distributed.

Another variation on the Wright-Panchapakesan statistic is also incorporated into BICAL as a measure of item fit. The between-t differs from the Wright-Panchapakesan statistic in that persons with different but adjacent scores may be included in the same score group. The between-t in BICAL is calculated using six score groups.

The second approach to the construction of a chi-square test of model fit is similar to the first in that a residual of observed and expected results is obtained, but differs in that the frequencies of both correct and incorrect responses are used in calculation of the statistic. That is, a 2 X J contingency table of response by score or ability group (where J is the number of groups) is set up and a chi-square statistic constructed from the difference between observed and expected frequencies in each cell.

Van den Wollenberg (1979) presents a statistic called Q_1 which is derived in this way. Respondents are first grouped by total score into k-1 groups (where k is the number of items). For each item, observed and expected frequencies of correct and incorrect responses are calculated for each score group and a chi-square statistic on (k-2) degrees of freedom is obtained. The chi-squares for all items are then summed to produce an overall test statistic which is distributed as a chi-square on (k-1)(k-2) degrees of freedom.

In calculating the expected frequencies for each cell, van den Wollenberg uses parameter estimates obtained under the conditional maximum likelihood estimation procedure. Since the derivation of Q_1 does not appear to depend on this approach to parameter estimation, the statistic should be usable under other estimation procedures.

Yen (1981) employs a similar approach to produce a statistic which she also calls Q_1. Respondents are arranged into ten groups on the basis of ability estimate, yielding a statistic which is appropriate for the two- and three-parameter models as well as the one-parameter model.

The Q_1 statistic proposed by both van den Wollenberg and Yen is given by the formula

$$Q_1 = \sum_{i=1}^{K-1} \frac{N_i(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})} + \sum_{j=1}^{K-1} \frac{N_i((1 - O_{ij}) - (1 - E_{ij}))^2}{E_{ij}(1 - E_{ij})}$$

where O_{ij} is the observed proportion of examinees in group i

who answer item j correctly

and N_i is the number of examinees in group i.

The third approach to the construction of chi-square tests of model fit uses the item response model property of invariance of parameter estimates when the model fits the data. That is, the likelihood ratio test compares the likelihood of the observed data when the parameter estimates are obtained for the group as a whole to the likelihood when parameter estimates are based on subgroups of the data. Andersen (1973) presents a statistic which is calculated as follows:

$$Z = -2 \log \left(L(\hat{\underline{b}}) / \prod_{r=1}^{k-1} L_r(\hat{\underline{b}}_r)\right)$$

where $L(\hat{\underline{b}})$ is the likelihood of the total sample data given the difficulty parameter estimates $\underline{b}$ for the k items.

and $L_r(\hat{\underline{b}}_r)$ is the likelihood in the subgroup with the score r given the estimates $\underline{b}_r$ calculated in that subgroup.

When the data fit the one-parameter model, it is expected that $\hat{\underline{b}}_r \rightarrow \hat{\underline{b}}$, but if deviation from the model occurs, $\hat{\underline{b}}_r$ will differ from $\hat{\underline{b}}$ in at least some of the score groups. In this case, the likelihood of the observed results when difficulty estimates are allowed to differ across subgroups will generally be greater than the likelihood given estimates based on overall results. Therefore deviation from the model will lead to values of the ratio less than one. Andersen (1973) proves that Z tends to a limiting chi-square distribution with (k-1)(k-2) degrees of freedom when $n_r \rightarrow \infty$. When group sizes are small, however, the approximation to the chi-square distribution will be poor. Wainer, Morgan, and Gustaffson (1980) suggest that 50 to 100 persons are needed in each score group. When there are fewer persons in each group, parameter estimates are likely to be unstable (Gustaffson, 1977). Andersen advises that adjacent score groups be pooled when sample sizes are small. The statistic will then have an approximate chi-square distribution with (k-1)(m-1) degrees of freedom where m is the number of score groups. Partitioning can also be performed on the basis of sex or some other variable (e.g., see Gustaffson, 1977). A likelihood ratio test has also been developed for the two-parameter normal ogive model (Bock & Lieberman, 1970).

The likelihood ratio test can only be used when maximum likelihood parameter estimates are obtained. Thus it is not available for use with programs such as BICAL or LOGIST, which employ different methods of estimation. Although the likelihood ratio test is asymptotically distributed as a chi-square, it suffers the same sample size problems as do the Wright-Panchapakesan statistic and its variations. Use of any of these statistics therefore requires caution.

Given the theoretical problems associated with the fit statistics described above, Monte Carlo studies can provide valuable information about the conditions, if any, under which the statistics are sensitive to violations of the model assumptions. Surprisingly, few Monte Carlo studies have been reported in the literature. Rogers and Hattie (in press) conducted an extensive simulation study to examine the behavior of the person and item fit statistics used in BICAL. For 500 persons and 15 items, Rogers and Hattie generated data to fit the one-, two-, and three-parameter models and to reflect two levels of multidimensionality. With 75 replications of each dataset, Rogers and Hattie found that the total-t statistic was unable to detect to any practical degree overall model misfit as a result of guessing, heterogeneity in discrimination parameters, or multidimensionality. The between-t statistic, which resembles the Wright-Panchapakesan statistic, showed some sensitivity to variation in discrimination parameters and to guessing, but was insensitive to multidimensionality. The statistic is probably best used as a marker for items which should be more closely examined or as an indicator of the comparative fit of the one-parameter model against the two- or three-parameter models.

Van den Wollenberg (1979), in an evaluation of the Q_1 statistic, found that it appeared to have a chi-square distribution with the degrees of freedom claimed. Van den Wollenberg generated ten replications of one-parameter model data for 1000 persons and test lengths of four to eight items. The approximation of the distribution of Q_1 to the chi-square distribution was tested by means of the Kolmogorov-Smirnov test of fit. For tests with difficulty parameters in the range (-2, 2), the chi-square approximation was adequate for all test lengths. For tests with extreme item parameters, in the range (-4, 4), the approximation was not as good. This disturbance reflects the sensitivity of Q_1, like all chi-square statistics, to small cell sizes.

Yen (1981) examined the behavior of the Q_1 statistic she derived using simulated data for 1000 persons and 36 items. Yen generated data to fit the one-, two- and three-parameter models and fitted one-, two- and three-parameter models to each dataset. Pearsonian chi-square goodness-of-fit tests indicated that when generating and estimating model were in agreement, the chi-square approximation to the distribution of Q_1 was reasonable, although the mean of the statistic was consistently higher than expectation. When the estimating model fitted fewer parameters than were used to generate the data, the distribution of Q_1 was no longer chi-square; the mean value of Q_1 increased substantially beyond that expected, except in the case where the two-parameter model was fitted to three-parameter data.

Gustaffson (1980) conducted simulations to study the behavior of the Andersen likelihood ratio statistic. Data were generated for test of length 15 and 30 items and for sample sizes of 150 and 300. Variation in discrimination of parameters was small (0.8, 1.0, 1.2) or large (0.5, 1.0, 1.5) and guessing parameters were either all zero or all .2. Gustaffson concluded from his results that sample sizes of 500-1000 and test lengths of 20-40 items are necessary to provide reasonable power in the likelihood ratio statistic for the detection of heterogeneity in discrimination parameters and guessing.

Summary

The statistical tests of model fit described here (and summarized in Table 1) do appear to have some value. Because they are sensitive to sample size and because they are not uniformly powerful, however, the use of any of these statistics as the sole indicator of model fit is clearly inadvisable. Use of the BICAL fit statistics to discard persons and items from model calibration procedures should certainly be warned against. The van den Wollenberg and Yen statistics need further evaluation, but appear to have promise, at least as contributors to the evidence of model fit. Until conditional maximum likelihood estimation procedures are incorporated into computer programs in this country, the likelihood ratio statistic will not see much use.

Although the weaknesses of chi-square tests of model fit must always be borne in mind, two situations can be identified in which these tests may lead to relatively clear interpretations. When sample sizes are small and the statistics indicate model misfit, or when sample sizes are large and model fit is obtained, the researcher may have reasonable confidence that, in the first case, the model does misfit the data, and in the second, that it fits the data. These possibilities make it worthwhile to employ statistical tests of fit despite their problems and despite the alternate possibility of equivocal results.

Assessing Item Dimensionality

The assumption that a set of test items is "unidimensional" is made for all of the presently popular item response models. Despite the importance of the assumption, there is substantial confusion in the psychometric literature concerning the proper definition of the term "unidimensionality" and methods for assessing its presence or absence in a set of test items (Hattie, 1984, 1985; Traub & Wolfe, 1981). Hattie (1984) reported that there are 87 indices in the psychometric literature for assessing the dimensionality of a set of test items. Unfortunately, these methods (or indices) are only loosely connected to the many definitions in the literature.

In some of our recent research (Hambleton & Rovinelli, in press), interest was centered on three promising methods for addressing the unidimensionality of a set of test items: (1) non-linear factor analysis (NLFA), (2) residual analysis, and (3) the Bejar analysis. The first method appeared promising because NLFA does not require the implausible assumption of linear relationships among the variables and between the variables and the underlying traits to be made. In fact, one of the fundamental assumptions of IRT is that these relationships are non-linear (Lord, 1980). The second method was an assessment of the overall fit of a unidimensional model to a dataset through the analysis of residuals. When the fit is adequate, it would seem that the assumption of a unidimensional model is plausible also (see, for example, Rentz & Rentz, 1979). The Bejar (1980) method appeared useful for assessing item dimensionality because it does not involve questionable linearity assumptions about the test data. In addition, the method provides a straightforward check on one of the expected outcomes of a unidimensional set of test data: the subset of items from a test in which an item is calibrated is irrelevant.

The specific purpose of the Hambleton and Rovinelli investigation was to compare the assessments of the dimensionality of a set of test items with four methods: linear factor analysis (LFA) , non-linear factor analysis, residual analysis, and Bejar analysis. The four methods were applied to five datasets. The datasets were artificial and generated to reflect one- and two-dimensional datasets.

Description of Methods

LFA is probably the most commonly used method for studying item dimensionality. Using (1) the matrix of phi or tetrachoric correlations to summarize the linear relationships between pairs of items in a test, and (2) communality estimates (often, squared multiple correlations) in the diagonal entries of the correlation matrix, eigenvalues are extracted from the correlation matrix and plotted (from largest to smallest). The number of "significant" factors is determined by looking for the "elbow" in the plot.

In NLFA, non-linear relationships between the variables and the traits or factors measured by the variables are assumed (McDonald, 1967). The application of NLFA to the study of item dimensionality seems especially desirable, within the context of item response theory, because one of the principal assumptions (i.e., the mathematical form of the item characteristic curves) specifies a particular non-linear relationship between item performance and ability.

The method for addressing the unidimensionality of a set of test items through a residual analysis involves fitting a unidimensional item response model of interest to the test data, using the model parameter estimates to predict the item performance data, and then summarizing the discrepancies or residuals (see, for example, Hambleton & Swaminathan, 1985). Specifically, ability categories are chosen to divide the ability scale into equal intervals. Examinees are assigned to categories based upon their ability estimates. For examinees in each ability category on each item, a comparison is made between actual performance (proportion correct) and the predicted proportion-correct level from the corresponding item characteristic curve (icc). The difference between the actual and predicted proportion-correct score (called a residual or a raw residual score) in each ability category and for each item can also be divided by the corresponding standard error of the proportion-correct estimate to obtain a standardized residual. When the chosen model fits the dataset, these standardized residuals might be expected to be small and randomly distributed about the value 0. The rationale for the appropriateness of residuals as a check on item unidimensionality is that when a unidimensional model fits a dataset, all of the model assumptions must be met to a reasonable degree.

Bejar (1980) argued that if the set of items in a test is unidimensional, then the grouping of test items from the test for the purpose of item calibration will be irrelevant. Parameter estimates for items calibrated with different subsets of items, aside from sampling errors, should be identical. Bejar's method (with minor modifications) can be implemented in four steps:

1. Identify a subset of items in the test which appears to be measuring a trait different from the trait measured by the total test.
2. Conduct a three-parameter model analysis of only the items in the subtest.
3. Repeat the three-parameter model analysis using the total set of items.
4. Compare the two sets of *b*-value estimates for items in the subtest.

The pairs of parameter estimates for items in the subtest and test, respectively, should be linearly related unless the subset of items is measuring a trait or traits which are not common to the trait or traits measured in the total test.

Description of the Data

To compare the four methods, Hambleton and Rovinelli (in press) generated five artificial test datasets were generated to be consistent with the assumption of either a one- or a two-dimensional latent space. Each test consisted of 40 test items. The item performance for 1500 examinees was simulated with the three-parameter logistic model. In dataset 1, the latent space was chosen to be one-dimensional. In datasets 2 to 5, the latent space was chosen to be two-dimensional. The only difference between datasets 2 and 3, and 4 and 5 was that in datasets 2 and 3 the correlation between the two latent traits was .10 whereas in datasets 4 and 5 the correlation between the two traits was .60. In addition, items were generated to measure one trait or the other. In datasets 2 and 4, the first 20 items measured trait one and the second 20 items measured trait two. In datasets 3 and 5, the first 30 items measured trait one and the remaining 10 items measured trait two. The chart below summarizes the pertinent information:

			Number of Items	
Dataset	Trait(s)	$r(\theta_1,\theta_2)$	First Trait	Second Trait
1	1	--	40	0
2	2	.10	20	20
3	2	.10	30	10
4	2	.60	20	20
5	2	.60	30	10

Parameter values were assigned to items on each trait in the following way:

_b_ parameters were drawn at random from a uniform distribution on the interval [-2.0, +2.0]

_a_ parameters were drawn at random from a uniform distribution on the interval [.40, 2.00]

_c_ parameters were set tu a value of .25.

The choice of item parameters reflected values often found in practice (Hambleton & Swaminathan, 1985).

Results - One-Dimensional Data

The results of fitting from one to five linear factors, and one and two factors with linear, linear and quadratic, and linear, quadratic, and cubic terms to the one-dimensional dataset are reported in Table 2. The first two criteria (r_{ij}, $s(r_{ij})$) show simply that the mean off-diagonal elements after fitting one or more factors are centered close to .00 (as compared to .127 in the original correlation matrix) and that the standard deviation of the distribution of the residuals approaches zero as the number of factors is increased. From the statistics in the third and fourth columns of Table 2 it is clear that a NLFA with one factor with linear and quadratic terms fits the data better than the two factor solution provided by LFA. In fact, even three linear factors did not produce as accurate a fit to the data.

The residual analyses for the one-dimensional data with the three logistic models are reported in Table 3. Not surprisingly, since the data were generated to fit the three-parameter model, this model provided the best fit to the data. More importantly, the distribution of standardized residuals (SRs) was (approximately) normal and the mean absolute-valued SR was close to .799. With the one-dimensional data and when the particular IRT model closely fits the data, the SRs appear to have the desired distribution.

Since, for this dataset, all of the test items were generated to fit a one-dimensional model, there was no reason to suspect that a second trait was necessary to account for the inter-item correlations. As a rather simple check on the method, the last 20 items were presumed to measure a second trait and the Bejar method applied. The correlation between the *b*-values was in excess of .99. Clearly, the assumption of unidimensionality could not be rejected on the basis of the available evidence, nor should the assumption be rejected for this dataset.

Results - Two-Dimensional Data

If the largest eigenvalue of the random data (=1.48) is used as the criterion for determining the number of factors for all four two-dimensional datasets, three significant factors emerged (see Table 2). Again, the linear factor analysis method resulted in more factors than the underlying dimensionality of the data.

Again, Table 2 shows that the NLFA method produced good results. With the two-dimensional data (r=.10; 20/20) and the two-dimensional data (r=.10; 30/10), the mean and standard deviation of absolute-valued residuals associated with a two-factor model with quadratic terms were smaller than those of the corresponding residuals obtained from a three-factor solution using LFA. Thus, if the three-factor solution with LFA is acceptable, then the two-factor solution from NLFA will be, too. The two-factor model with cubic terms was not obtained because of the high costs associated with running the computer program and the acceptability of the two-factor solution with quadratic terms.

Table 3 provides a summary of the absolute-valued residuals and standardized residuals obtained from fitting logistic models to the four two-dimensional datasets. Several findings are evident:

1. The one-parameter model did not fit any of the datasets. Rather than suggesting multidimensionality in the data, however, the likely explanation in view of the results of fitting the one-parameter model to the one-dimensional data is that the misfit is due to the failure of the model to account for variations in item discrimination power and the guessing behavior of low-ability examinees.
2. A comparison of the SRs from the two- and three-parameter models showed substantially smaller SRs than those obtained with the one-parameter model, and the three-parameter model fitting the datasets slightly better than the two-parameter model. On the basis of a study of the SRs for the two- and three-parameter models, a researcher would probably accept the hypothesis that the test items in each dataset were unidimensional.
3. The overall fits were better when the traits were correlated (r=.60), than when the traits were not (r=.10).

How can the three-parameter model fit the four two-dimensional datasets? The failure to identify multidimensionality in datasets 4 and 5 was surprising but in view of the moderately high correlation between the two traits the results were not totally unexpected. LOGIST simp!y proceeds to estimate the second order factor which incorporates the two related factors. Why multidimensionality could not be detected in datasets 2 and 3 is not completely clear. This result was very disappointing. It appears that LOGIST estimates an average ability of the two unrelated traits and also attaches low a-values to all of the test items. In doing so, a reasonable fit between the model and each dataset can be achieved. When there is an imbalance in the test data (i.e., 30/10), LOGIST assigns high a-values to items measuring the "dominant trait" and relatively low values to the remaining items. In this way, a one-dimensional model can fit the data. With a more even split (i.e., 20/20), the values assigned to the a-values are relatively low. In any case, because of the way LOGIST handles multidimensionality in the test data, residual analyses cannot identify it when it is present.

The results of the Bejar analyses on the four two-dimensional datasets were especially surprising as well as disappointing:

1. With r=.10, and a split of 20/20, the test items had comparable b-values.
2. With r=.10, and a split of 30/10, the b-values were substantially different and appeared to be poorly estimated. This analysis would lead to a rejection of the unidimensionality assumption.

3. When r=.60, and for the two splits 20/20 and 30/10, the Bejar analyses suggested that the assumption of unidimensionality could not be rejected.

In only one of the four analyses was the Bejar method sensitive to the multidimensionality in the data.

Though the results were not reported in detail by Hambleton and Rovinelli, the four methods for assessing dimensionality were also applied to the 80 item section of the 1982 ABFP In-Training Exam. The four methods provided different answers to the question of unidimensionality! Had the simulation studies described earlier not been carried out, the results from the residual analyses or the Bejar analyses would have been used to support the assumption of unidimensionality. The LFA of the data suggested that anywhere from 4 to as many as 8 significant factors would need to be retained for a satisfactory accounting of the data. The NLFA also appears to indicate that more than one factor may be needed. In summary, the four methods provided contradictory information about the item dimensionality. Based upon the results from the simulations, it would seem that the most likely conclusion is that more than one dimension is operating.

Conclusions

On the basis of a single simulation study with limited scope, generalizability of the findings is obviously limited. But several findings of the study do appear to suggest directions for some future work. First, the linear factor analysis model in all instances overestimated the number of underlying dimensions in the data. Second, non-linear factor analysis with linear and quadratic terms led to the correct determination of the item dimensionality in the three datasets where it was used.

Both the residual analysis method and Bejar's method provided disappointing results . It appears that the three-parameter model can accommodate multidimensionality by assigning low *a*-values to these "deviant" items. Good fit is achieved, but in doing so, the "deviant" items are essentially removed from the test since those items neither contribute much to ability parameter estimation or to the test information function. Likewise, the Bejar method was unable to detect the two underlying traits except when the correlation between the traits was low and a disproportionate number of the test items measured one of the traits.

In conclusion, despite the limited scope of the present investigation, the results do suggest the need for extreme caution in using linear factor analysis, residual analysis, or Bejar's method to address questions about item unidimensionality. Clearly, more investigations of these methods showing some positive results are needed before they can be strongly recommended for use by practitioners. On the other hand, while non-linear factor analysis produced the most promising results in this study, an accepted criterion for determining the minimum number of factors to retain in a non-linear factor solution is not available, nor is an easy-to-use non-linear factor analysis program available. More research along these lines must be carried out first before NLFA can be recommended.

Studies of Item and Ability Parameter Invariance

Regardless of the extent to which a set of model assumptions are met by a set of test data, researchers commonly proceed to check model data fit and (occasionally) the invariance property of the item and ability parameter estimates. The rationale seems to be that (1) little is known about model robustness and (2) the tests of model assumptions are not well-developed or grounded in statistical theory, and therefore the additional checks on model appropriateness are justified. Studies with respect to the invariance property for the most part have been viewed as item bias investigations: studies which check the extent to which item parameter estimates determined from different subgroups of a population are equivalent. In a few cases, researchers have checked the invariance property of ability estimates by comparing ability estimates obtained from hard and easy sets of items in a test (e.g. Wright, 1968).

Our own research has centered on the plot method for checking the presence of the invariance property (Hambleton & Murray, 1983; Hambleton & Rogers, 1986; Hambleton, Rogers, & Arrasmith, 1986). In this method, an independent variable of interest is chosen (e.g. sex). Then, two groups of (say) males and females are formed and then the groups are divided again to form two randomly equivalent female samples and two randomly equivalent male samples. Three-parameter model analyses are conducted on each of the four groups. Finally, the plots of the item statistics (especially the b-values) in the randomly-equivalent groups are compared to the plots of the b-values in the female and male samples. When sex is not a factor in item performance, all the plots should be similar in shape. If the plots differ, sex is assumed to be a factor and items showing consistently large differences in their item statistics are identified for further review. The plot method has several advantages: (1) provides a graphical solution to the item invariance problem that is easy to understand, and (2) recognizes the instability in item bias statistics by focusing only on items which show consistently large differences across independent samples.

One problem that arises is choosing a cut-off score for identifying items showing consistently large differences. The same problem arises with other item bias methods too, such as the "Total Area Method" and the "Root Mean Squared Difference Method." With the plot method, the distribution of (say) b-value differences between the randomly equivalent samples can serve as the sampling distribution under the null-hypothesis that there are no differences. This distribution can be used to determine the cut-off points corresponding to 1% and 5% type I errors. Then the cut-off points are applied to distribution of b-value differences in the male and female samples so that a subset of potentially biased items can be identified. Unfortunately, the distribution of b-value differences in the plot method is achieved at a cost: sample sizes must be cut in half and this action has a negative influence on the precision of item parameter estimates. As an alternative, in some recent work, we have been using sampling distributions of b-value differences, total area statistics, and root mean squared difference statistics generated through item response models from randomly-equivalent samples of simulated data. Figures and 2 and 3 highlight this work with the total area statistic. Figure 2 highlights the similarity in the distributions of total area statistics from randomly-equivalent groups using real and simulated data. The results are very similar. Figure 3(a) shows a distribution of total area statistics for 75 items obtained from real data for male and female samples. Figure 3(b) as well as Figures 2(a) and (b) can be used for the purposing of setting cut-off points. The important point is the high similarity in the distributions generated under the null hypothesis for real data in Figure 2(a) with simulated data in Figure 2(a) and 3(b). These results along with others we have, strongly supported the use of logistic models to generate data for the purpose of producing base-line statistics for interpreting invariance studies.

Conclusions

The potential of item response theory has been widely documented but that potential is not necessarily assured. Choice of tests, populations of examinees, and applications will influence the success of IRT uses. With respect to addressing the fit between an item response model and a set of test data for some desired application, our view at this time is that the best approach involves (1) designing and implementing a wide variety of analyses, (2) interpreting the full set of results carefully, and (3) judgmentally determining the appropriateness of the intended application. Table 5 provides some initial thoughts on appropriate goodness-of-fit investigations for both small and large scale investigations.

Analyses should include investigations of model assumptions, the extent to which desired model features are obtained, and comparisons between model predictions and actual data. Statistical tests can be carried out but care must be shown in interpreting the statistical information. Model misfit with small samples or satisfactory fit obtained with large samples are especially useful results with statistical tests. Extensive use should be made of replication, cross-validation, and of graphical displays of model predictions and actual data, etc. Also, fitting more than one model and comparing the residuals provides information that is invaluable in determining the usefulness of item response models. Whenever possible, investigate the consequences of misfit. There is no limit to the number of investigations that can be carried out. The amount of effort in collecting, analyzing, and interpreting results should be related to the importance and nature of the intended application. For example, small school districts using the one-parameter model in item banking and test development for classroom tests will not need to expend as many resources'on goodness-of-fit studies as, say, the Educational Testing Service when they equate multiple forms of nationally standardized aptitude tests.

In summary, our recommended strategy for assessing model-data fit is to accumulate a considerable amount of evidence that can be used to aid in the determination of the appropriateness of a particular use of an item response model. Judgment will ultimately be required and therefore the more evidence available, the more informed the final decision about the use of an item response model will be. In fact, goodness-of-fit studies and the interpretations of results are very much like studies to validate tests as measures of constructs. A researcher can never prove that a test measures a construct. However, he/she can accumulate enough evidence so that reasonable persons can agree that it makes sense to assume the test measures the construct until counter-evidence is available.

References

Andersen, E.B. (1973). A goodness-of-fit test for the Rasch model. Psychometrika, 38, 123-139.

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), Handbook of Methods for Detecting Test Bias. Baltimore, MD: The Johns Hopkins University Press.

Ansley T.N., & Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement, 9, 37-48.

Baker, F.B. (1964). An intersection of test score interpretation and item analysis. Journal of Educational Measurement, 1, 23-28.

Baker, F.B. (1965). Origins of the X_{50} and B as a modern item analysis technique. Journal of Educational Measurement, 2, 167-180.

Baker, F.B. (1986). Two parameter: The forgotten model. Paper presented at the annual meeting of NCME, San Francisco.

Bejar, I.I. (1980). A procedure of investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17, 283-296.

Bock, R.D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psycholetrika, 35, 179-197.

Cronbach, L.J., & Warrington, W.G. (1951). Time-limit tests: Estimating their reliability and degree of speeding. Psychometrika, 16, 167-188.

Divgi, D.R. (1981). Does the Rasch model really work? Not if you look closely. Paper presented at the annual meeting of NCME, Los Angeles.

Drasgow, F., & Lissak, R.I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. Journal of Applied Psychology, 68, 363-373.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology, 33, 234-246.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Gustaffson, J.E. (1980). Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 33, 205-233.

Hambleton, R.K. (1980). Latent ability scales, interpretations, and uses. In S. Mayo (Ed.), New directions for testing and measurement: Interpreting test scores (no. 6). San Francisco: Jossey-Bass.

Hambleton, R.K., & Cook, L.L. (1983). The robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. Weiss (Ed.), New Horizons in Testing. New York: Academic Press.

Hambleton, R.K., & Murray, L. (1983). Some goodness of fit investigations for item response models. In R.K. Hambleton (Ed.), Applications of Item Response Models. Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R.K., Murray, L., & Simon, R. (1982). Utilization of item response models with NAEP mathematics exercise results. Final Report (ECS Contract No. 02-81-20319). Submitted to the Educational Commission of the States and the National Institute of Education.

Hambleton, R.K., Rogers, H.J., & Arrasmith, D. (1986). A comparison of the Mantel-Haenzel Statistic and item response theory methods of identifying difference item performance. Paper presented at a symposium at the annual meetings of AERA and NCME, San Francisco.

Hambleton, R.K., & Rogers, H.J. (1986). Evaluation of the plot method for identifying potentially biased test items. In S.H. Irvine, S. Newstead, & P. Dann (Eds.), Computer-based human assessment. Hingham, MA: Kluwer-Nijhoff.

Hambleton, R.K., & Rovinelli, R. (1973). A FORTRAN IV program for generating examinee response data from logistic test models. Behavioral Science, 18, 74.

Hambleton, R.K., & Rovinelli, R.J. (in press). Assessing the dimensionality of a set of test items. Applied Psychological Measurement.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff Publishers.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. (1978). Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 467-510.

Hambleton, R.K., & Traub, R.E. (1973). Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 26, 195-211.

Hattie, J.A. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49-78.

Hattie, J. (1985). Methodological review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Hills, J.R., Beard, J.G., Yotinprasert, S., Roca, N.R., & Subhiya, R.G. (1985). An investigation of the feasibility of using the three-parameter model for Florida's Statewide Assessment Tests. Tallahassee, FL: College of Education, Florida State University.

Holland, P.W. (1981). When are item response models consistent with observed data? *Psychometrika, 46*, 79-82.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Kingston, N.M., & Dorans, N.J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147-154.

Kingston, N.M., & Dorans, N.J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9*, 281-288.

Linn, R.L., & Harnisch, D.L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*, 109-118.

Lord, F.M., (1970). Estimating item characteristic curves without knowledge of their mathematical form. *Psychometrika, 35*, 43-50.

Lord, F.M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247-264.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*, Hillsdale, NJ: Erlbaum.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McDonald, R.P. (1967). Non-linear factor analysis. *Psychometric Monographs*, No. 15.

McDonald, R.P. (1980a). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 33*, 205-233.

McDonald, R.P. (1980b). Fitting latent trait models. In D. Spearitt (Ed.), *The Improvement of Measurement in Education and Psychology*. Proceedings of the Invitational Seminar for the Fiftieth Anniversary of the Australian Council of Educational Research, Melbourne.

McKinley, R.L., & Mills, C.N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 9, 49-57.

Popham, W.J. (1980). Modern educational measurement. ~.iglewood Cliffs, NJ: Prentice-Hall.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Rentz, R.R., & Rentz, C.C. (1979). Does the Rasch model really work? NCME Measurement in Education, 10, 1-11.

Rogers, H.J., & Hattie, J.A. (in press). A Monte Carlo investigation of several person and item fit statistics for item response models. Applied Psychological Measurement.

Rosenbaum, P.R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika, 49, 425-435.

Traub, R.E., & Wolfe, R.G. (1981). Latent trait theories and the assessment of educational achievement. In D.C. Berliner (Ed.), Review of Research in Education - Volume 9. Washington, D.C.: American Educational Research Association.

Wainer, H., Morgan, A., & Gustaffson, J. (1980). A review of estimation procedures for the Rash model with an eye towards longish tests. Journal of Educational Statistics, 5, 35-64.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). Princeton, NJ: Educational Testing Service.

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

Wollenberg, A.L. van den (1982b). A simple and effective method to test the unidimensionality axiom of the Rasch model. Applied Psychological Measurement, 6, 83-91.

Wright, B.D. (1968). Sample free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems.

Wright, B.D., Mead, R.J., & Bell, S.R. (1979). BICAL: Calibrating items with the Rasch model. (Research Memorandum 23-B). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

AERA in 86, Bias.2

Wright, B.D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.

Wright, B.D., & Stone, M.H. (1979). Best Test Design. Chicago: MESA Press.

Yen, W.M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 1980, 17, 297-311.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5(2), 245-262.

Figure 1. Approaches for Conducting Goodness-of-Fit Investigations. (An update of Figure 8.1 which appeared in Hambleton & Swaminathan, 1985)

Checking Model Assumptions

1. Unidimensionality (Applies to Nearly All of the Popular Item Response Models)

   - Plot of Eigenvalues (from Largest to Smallest) of the Inter-Item Correlation Matrix (Tetrachoric Correlations Preferable to Phi Correlations)--Look for a dominant first factor, and a high ratio of the first to the second eigenvalue (Reckase, 1979).
   - Comparison of Two Plots of Eigenvalues--The one described above and a plot of eigenvalues from an inter-item correlation matrix of random data (same sample size, and number of variables, random data normally distributed) (Horn, 1965).
   - Plot of Content-Based Versus Total-Test-Based Item Parameter Estimates (Bejar, 1980).
   - Analysis of Residuals After Fitting a One-Factor Model to the Inter-Item Covariance Matrix (McDonald, 1980a, 1980b).
   - Non-Linear Factor Analysis with Analysis of Residuals (Hambleton & Rovinelli, in press).
   - Modified Parallel Analysis (Drasgow & Lissak, 1983).
   - A Test of Local Non-Negative Dependence (Holland, 1981) (rejection of this hypothesis implies rejection of all IRT models which assume local independence).

2. Equal Discrimination Indices (Applies to the One-Parameter Logistic Model)

   - Analysis of Variability of Item-Test Score Correlations (for Example, Point-Biserial and Biserial Correlations).
   - Identification of Percent of Item-Test Score Correlations Falling Outside Some Acceptable Range (for Example, the Average Item-Test Score Correlation ± .15).

Figure 1, continued:

3. Minimal Guessing (Applies to the One- and Two-Parameter Logistic Models)

   - Investigation of Item-Test Score Plots (Baker, 1964, 1965).
   - Consideration of the Performance of Low-Ability Examinees (Selected with the Use of Test Results, or Instructor Judgments on the Most Difficult Test Items)
   - Consideration of Item Format and Test Time Limits (for Example, Consider the Number of Item Distractors, and Whether or Not the Test Was Speeded).

4. Nonspeeded (Power) Test Administration (Applies to Nearly All Item Response Models).

   - Comparison of Variance of the Number of Items Omitted to the Variance of the Number of Items Answered Incorrectly (Gulliksen, 1950).
   - Investigation of the Relationship Between Scores on a Test with the Specified Time Limit and with an Unlimited Time Limit (Cronbach and Warrington, 1951).
   - Investigation of (1) Percent of Examinees Completing the Test, (2) Percent of Examinees Completing 75 Percent of the Test, and (3) Number of Items Completed by 80 Percent of the Examinees.

5. Mathematical Form of the ICCs.

   - Tests of Monotonocity of ICCs (Rosenbaum, 1984).

Checking Expected Model Features

1. Invariance of Item Parameter Estimates (Applies to All Models)

   - Comparison of Item Parameter Estimates Obtained in Two or More Subgroups of the Population for Whom the Test is Intended (for Example, Males and Females; Blacks, Whites, and Hispanics; Instructional Groups; High and Low Performers on the Test or Other Criterion Measure, Geographic Regions). Normally, comparisons are made of the item-difficulty estimates and presented in graphical form (scattergrams). Random splits of the population into subgroups the same size provide a basis for obtaining plots which can serve as a baseline for

Figure 1, continued:

interpreting the plots of principal interest (Angoff, 1982; Lord, 1980; Hambleton and Murray, 1983). Graphical displays of distributions of standardized differences in item parameter estimates can be studied. Distributions ought to have a mean of zero and a standard deviation of one (for example, Wright, 1968).

2. Invariance of Ability Parameter Estimates (Applies to All Models)

- Comparison of Ability Estimates Obtained in Two or More Item Samples from the Item Pool of Interest. Choose item samples which have special significance such as relatively hard versus relatively easy samples, and subsets reflecting different content categories within the total item pool. Again, graphical displays and investigation of the distribution of ability differences are revealing.

Checking Model Predictions of Actual (and Simulated) Test Results

- Investigation of Residuals and Standardized Residuals of Model-Test Data Fits at the Item and Person Levels. Various statistics are available to summarize the fit information. Graphical displays of data can be revealing.

- Comparison of Item Characteristic Curves Estimated in Substantially Different Ways (for Example, Lord, 1970).

- Plot of Test Scores and Ability Estimates (Lord, 1974).

- Plots of True and Estimated Item and Ability Parameters (for Example, Hambleton & Cook, 1983). These studies are carried out with computer simulation methods.

- Comparison of Observed and Predicted Score Distributions. Various statistics (chi-square, for example) and graphical methods can be used to report results. Cross-validation procedures should be used, especially if sample sizes are small (Hambleton & Traub, 1973).

- Investigation of Hypotheses Concerning Scoring Keys (Kingston & Dorans, 1985), Practice Effects, Test Speededness, Cheating, Boredom, Item Format Effects (Kingston & Dorans, 1985), Item Order (Kingston & Dorans, 1985), etc.

- Comparisons of Two-Dimensional Data with One Dimension of Model Results (for example, Ansley & Forsyth, 1985).

(a)

HISTOGRAM total area statistics (F_1-F_2 and M_1-M_2), Real Data

```
                                              SYMBOL  COUNT       MEAN      ST.DEV.
                                                X      150        .275         .'74
                                   EACH SYMBOL REPRESENTS          1 OBSERVATIONS
INTERVAL                                                     FREQUENCY PERCENTAGE
NAME            5    10    15    20    25    30    35      INT. CUM.  INT.  CUM.
           +----+----+----+----+----+----+----+
UPTO 0.0   +                                                    0    0    .0    .0
 0.0+0.1   +XXXXXXXXXXXXXXXXXXX                                19   19  12.7  12.7
+0.1+0.2   +XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX*               37   56  24.7  37.3
+0.2+0 3   +XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX*               41   97  27.3  64.7
+0.3+0.4   +XXXXXXXXXXXXXXXXXXXXX                              21  118  14.0  78.7
+0.4+0.5   +XXXXXXXXXXXXXXXXXXXX                               20  138  13.3  92.0
+0.5+0.6   +XXXXXXXX                                            8  146   5.3  97.3
+0.6+0.7   +X                                                   1  147    .7  98.0
+0.7+0.8   +X                                                   1  148    .7  98.7
+0.8+0.9   +X                                                   1  149    .7  99.3
+0.9+1.0   +                                                    0  149    .0  99.3
+1.0+1.1   +                                                    0  149    .0  99.3
+1.1+1.2   +                                                    0  149    .0  99.3
+1.2+1.3   +                                                    0  149    .0  99.3
+1.3+1.4   +X                                                   1  150    .7 100.0
+1.4+1.5   +                                                    0  150    .0 100.0
OVER+1.5   +                                                    0  150    .0 100.0
           +----+----+----+----+----+----+----+
                5    10    15    20    25    30    35
```

(b)

HISTOGRAM total area statistics (F_1-F_2 and M_1-M_2), Simulated Data

```
                                              SYMBOL  COUNT       MEAN      ST.DEV.
                                                X      150        .232         .170
                                   EACH SYMBOL REPRESENTS          1 OBSERVATIONS
INTERVAL                                                     FREQUENCY PERCENTAGE
NAME            5    10    15    20    25    30    35      INT. CUM.  INT.  CUM.
           +----+----+----+----+----+----+----+
UPTO 0.0   +                                                    0    0    .0    .0
 0.0+0.1   +XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX                  33   33  22.0  22.0
+0.1+0.2   +XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX*               47   80  31.3  53.3
+0.2+0.3   +XXXXXXXXXXXXXXXXXXXXXXXXXXXXX                      29  109  19.3  72.7
+0.3+0.4   +XXXXXXXXXXXXXXXXXXXX                               20  129  13.3  86.0
+0.4+0.5   +XXXXXXXXXXXX                                       12  141   8.0  94.0
+0.5+0.6   +XXXX                                                4  145   2.7  96.7
+0.6+0.7   +                                                    0  145    .0  96.7
+0.7+0.8   +XX                                                  2  147   1.3  98.0
+0.8+0.9   +XX                                                  2  149   1.3  99.3
+0.9+1.0   +X                                                   1  150    .7 100.0
+1.0+1.1   +                                                    0  150    .0 100.0
+1.1+1.2   +                                                    0  150    .0 100.0
+1.2+1.3   +                                                    0  150    .0 100.0
+1.3+1.4   +                                                    0  150    .0 100.0
+1.4+1.5   +                                                    0  150    .0 100.0
OVER+1.5   +                                                    0  150    .0 100.0
           +----+----+----+----+----+----+----+
                5    10    15    20    25    30    35
```

Figure 2. Histograms of total area item bias statistics calculated between randomly equivalent groups using reading competency test data in (a) and simulated data in (b).

(a)

HISTOGRAM total area statistics (F_1-M_1 and F_2-M_2), Real Data

```
                                          SYMBOL   COUNT      MEAN      ST.DEV.
                                            X       150        .305         .212
                               EACH SYMBOL REPRESENTS            1 OBSERVATIONS
INTERVAL                                               FREQUENCY PERCENTAGE
NAME            5    10    15    20    25    30    35  INT. CUM.  INT.  CUM.
           +----+----+----+----+----+----+----+
UPTO 0.0   +                                            0     0     .0    .0
 0.0+0.1   +XXXXXXXXXXXXXXXXXXXXX                      21    21   14.0  14.0
+0.1+0.2   +XXXXXXXXXXXXXXXXXXXXXXXXXXXXX              29    50   19.3  33.3
+0.2+0.3   +XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX*        37    87   24.7  58.0
+0.3+0.4   +XXXXXXXXXXXXXXXXXXXXXXXXX                  25   112   16.7  74.7
+0.4+0.5   +XXXXXXXXXXXXXXX                            15   127   10.0  84.7
+0.5+0.6   +XXXXXXXXXXXX                               12   139    8.0  92.7
+0.6+0.7   +XXXX                                        4   143    2.7  95.3
+0.7+0.8   +XXX                                         3   146    2.0  97.3
+0.8+0.9   +X                                           1   147     .7  98.0
+0.9+1.0   +XX                                          2   149    1.3  99.3
+1.0+1.1   +                                            0   149     .0  99.3
+1.1+1.2   +                                            0   149     .0  99.3
+1.2+1.3   +                                            0   149     .0  99.3
+1.3+1.4   +                                            0   149     .0  99.3
+1.4+1.5   +X                                           1   150     .7 100.0
OVER+1.5   +                                            0   150     .0 100.0
           +----+----+----+----+----+----+----+
                5    10    15    20    25    30    35
```

(b)

HISTOGRAM total area statistics (F_1-M_1 and F_2-M_2), Simulated Data

```
                                          SYMBOL   COUNT      MEAN      ST.DEV.
                                            X       150        .250         .214
                               EACH SYMBOL REPRESENTS            1 OBSERVATIONS
INTERVAL                                               FREQUENCY PERCENTAGE
NAME            5    10    15    20    25    30    35  INT. CUM.  INT.  CUM.
           +----+----+----+----+----+----+----+
UPTO 0.0   +                                            0     0     .0    .0
 0.0+0.1   +XXXXXXXXXXXXXXXXXXXXXXXXX                  25    25   16.7  16.7
+0.1+0.2   +XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX*        49    74   32.7  49.3
+0.2+0.3   +XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX*        37   111   24.7  74.0
+0.3+0.4   +XXXXXXXXXXXXXXXX' XX                       18   129   12.0  86.0
+0.4+0.5   +XXXXXXXXXX                                 10   139    6.7  92.7
+0.5+0.6   +XXX                                         3   142    2.0  94.7
+0.6+0.7   +XXX                                         3   145    2.0  96.7
+0.7+0.8   +XXX                                         3   148    2.0  98.7
+0.8+0.9   +                                            0   148     .0  98.7
+0.9+1.0   +                                            0   148     .0  98.7
+1.0+1.1   +X                                           1   149     .7  99.3
+1.1+1.2   +                                            0   149     .0  99.3
+1.2+1.3   +                                            0   149     .0  99.3
+1.3+1.4   +                                            0   149     .0  99.3
+1.4+1.5   +                                            0   149     .0  99.3
OVER+1.5   +X                                           1   150     .7 100.0
           +----+----+----+----+----+----+----+
                5    10    15    20    25    30    35
```

Figure 3. Histograms of total area item statistics calculated for female and male samples using reading competency test data in (a) and simulated data in (b).

Table 1
Summary of x^2 Statistics Used to Test Model Fit

Statistic	Citation	Distinguishing Features
Wright-Panchapakesan	Wright & Panchapakesan (1969)	- examinees grouped by total score into (k-1) groups, k = no. of items - standardized residual of observed and expected frequencies calculated within groups - residual squared and summed over groups and items for overall fit - $x^2(k-1)(k-2)$
Mean Square Residual	Wright & Stone (1979)	- based on individual observations - standardized residual of observed response (1/0) and probability of correct response calculated - residual squared and summed ove, persons or items for item or person fit - transformed to normal distribution
Total-t	Wright, Mead, & Bell (1979)	- Incorporated in BICAL - ratio of variance estimates - based on individual observations - summed over persons or items for item or person fit - transformed to a normal distribution in BICAL
Between-t	Wright, Mead, & Bell (1979)	- Incorporated in BICAL - variation on Wright-Panchapakesan statistic - uses six score groups - summed over groups for item fit - $x^2(5)$

-continued on next page-

Table 1, continued:

Statistic	Citation	Distinguishing Features
van den Wollenberg Q_1	van den Wollenberg (1979)	- based on contingency tables of response (1/0) by score level (1, ..., k-1) for each item - difference between observed and expected frequencies calculated for each cell to give x^2 on (k-2) degrees of freedom - expected frequencies calculated using conditional maximum likelihood parameter estimates - summed over all items for overall fit - x^2(k-1)(k-2)
Yen Q_1	Yen (1981)	- similar in construction to van den Wollenberg Q_1 - appropriate for one-, two-, and three-parameter models - 10 groups formed using ability estimate - x^2(10-s); s= no. of item parameters in model
Andersen Z	Andersen (1973)	- likelihood ratio statistic - uses conditional maximum likelihood parameter estimates - ratio of likelihood using parameter estimates obtained for total sample to product of likelihoods for different subgroups - x^2(k-1)(k-2)

Table 2
Eigenvalues (λ) and Percent of Variance Accounted for in
Random and One-Dimensional Datasets Using Phi
and Tetrachoric Correlations
(40 items; 1500 examinees)

	Random Data[1]		One-Dimensional Data			
	Tetrachoric		Phi		Tetrachoric	
Factor	λ	%	λ	%	λ	%
1	1.48	3.7	8.86	22.2	15.00	37.5
2	1.44	3.6	2.09	5.2	2.21	5.5
3	1.37	3.4	1.11	2.8	1.15	2.9
4	1.34	3.3	1.05	2.6	1.10	2.7
5	1.32	3.3	1.03	2.6	1.08	2.7
6	1.30		1.02		1.02	
7	1.28		.98		.96	
8	1.25		.96		.95	
9	1.22		.95		.92	
10	1.21		.93		.90	
11	1.19		.93		.88	
12	1.18		.93		.84	
13	1.15		.91		.80	
14	1.13		.89		.77	
15	1.10		.86		.74	

[1] Squared multiple correlations used as communality estimates.

Table 3
Residual Matrices After Fitting Linear
and Non-Linear Factor Analysis Models

				Goodness of Fit			
Dataset	Model	λ_1	% Var.	$\overline{r_{ij}}$	$s(r_{ij})$	$\overline{\|r_{ij}\|}$	$s(\|r_{ij}\|)$
1-DIM	Correlation Matrix			.127	.079	.127	.079
	Factor Analysis						
	1 Factor	6.64	16.6	.006	.078	.060	.050
	2 Factors	1.84	4.6	-.002	.030	.022	.021
	3 Factors	1.13	2.8	-.003	.025	.019	.016
	4 Factors	1.11	2.8	.000	.021	.016	.013
	5 Factors	1.10	2.7	.000	.019	.015	.012
	Non-Linear Factor Analysis						
	1 Factor, Linear Term			.002	.033	.026	.021
	1 Factor, Quad Term			.001	.022	.017	.014
	1 Factor, Cubic Term			.000	.022	.017	.014
	2 Factors, Linear Terms			-.006	.030	.022	.020
	2 Factors, Quad Terms			.000	.020	.015	.012
2-DIM	Correlation Matrix			.075	.090	.081	.084
(r=10;	Factor Analysis						
20/20)	1 Factor	4.41	11.0	.016	.074	.054	.054
	2 Factors	3.59	9.0	.000	.033	.024	.022
	3 Factors	1.64	4.1	.000	.025	.019	.016
	4 Factors	1.41	3.5	.000	.020	.016	.012
	5 Factors	1.15	2.9	.000	.018	.014	.011
	Non-Linear Factor Analysis						
	1 Factor, Linear Term			.025	.072	.050	.057
	1 Factor, Quad Term			.011	.037	.028	.027
	1 Factor, Cubic Term			.007	.029	.022	.020
	2 Factors, Linear Terms			-.005	.039	.027	.028
	2 Factors, Quad Terms			.000	.020	.016	.012

-continued on next page-

Table 3, continued

				Goodness of Fit			
Dataset	Model	λ_1	% Var.	$\bar{r}_{ij}$	$s(r_{ij})$	$\overline{\lvert r_{ij}\rvert}$	$s(\lvert r_{ij}\rvert)$
2-DIM (r=.10; 30/10)	Correlation Matrix			.104	.100	.109	.095
	Factor Analysis						
	1 Factor	6.27	15.7	.004	.047	.033	.034
	2 Factors	2.10	5.3	.002	.036	.029	.021
	3 Factors	1.88	4.7	.000	.023	.017	.015
	4 Factors	1.28	3.2	.000	.020	.016	.012
	5 Factors	1.09	2.7	.000	.018	.015	.011
	Non-Linear Factor Analysis						
	1 Factor, Linear Term			.008	.046	.033	.034
	1 Factor, Quad Term			.007	.036	.032	.029
	1 Factor, Cubic Term			.005	.039	.027	.028
	2 Factors, Linear Terms			-.004	.042	.031	.028
	2 Factors, Quad Terms			.000	.020	.016	.012
2-DIM (r=.60; 20/20)	Correlation Matrix			.111	.069	.111	.068
	Factor Analysis						
	1 Factor	5.7	14.3	-.001	.046	.038	.028
	2 Factors	2.2	5.6	.000	.030	.022	.020
	3 Factors	1.6	3.9	.000	.023	.018	.014
	4 Factors	1.2	3.1	.000	.020	.016	.013
	5 Factors	1.1	2.8	.000	.019	.015	.012
2-DIM (r=.60; 30/10)	Correlation Matrix			.132	.080	.132	.080
	Factor Analysis						
	1 Factor	6.8	16.9	.000	.042	.032	.027
	2 Factors	2.0	5.1	.000	.028	.021	.019
	3 Factors	1.6	3.9	.000	.021	.017	.013
	4 Factors	1.2	3.1	.000	.020	.015	.012
	5 Factors	1.1	.7	.000	.028	.014	.011

Table 4
Summary of Standardized Residuals (SRs)

Data Set and Model	% of Absolute-Valued SRs				Average Absolute-Valued Residual	Average Absolute-Valued SR
	|0 to 1|	|1 to 2|	|2 to 3|	|3 and over|		
1-DIM						
1	32.3	28.2	18.8	21.4	.067	1.86
2	66.6	26.8	5.5	1.1	.033	.89
3	76.8	21.1	1.8	.2	.031	.71
2-DIM (r=.10; 20/20)						
1	49.8	32.7	13.0	4.6	.048	1.20
2	63.6	32.1	3.0	1.4	.036	.86
3	68.2	26.8	3.9	1.1	.035	.84
2-DIM (r=.10; 30/10)						
1	33.2	26.6	16.4	23.9	.075	1.99
2	61.8	27.3	7.7	3.2	.038	.99
3	69.8	26.6	3.6	0.0	.027	.76
2-DIM (r=.60; 20/20)						
1	44.3	26.8	16.6	12.3	.060	1.51
2	67.1	24.8	7.1	1.1	.035	.88
3	72.7	22.7	3.6	0.9	.030	.79
2-DIM (r=.60; 30/10)						
1	39.1	25.7	15.0	20.2	.065	1.79
2	61.6	29.1	5.0	4.3	.038	1.00
3	73.2	24.1	2.7	0.0	.026	.73

Table 5
Suggestions for Model Selection/Goodness of Fit Investigations

Category	Small-Scale Applications (e.g., Classroom Tests)	Large-Scale Applications (Major State and Nationally Administered Tests)
Model Assumptions	- Conduct a classical item analysis (if the test has many hard items, and/or wide range of item discrimination indices, avoid the one-parameter model). - Avoid the three-parameter model with small sample. .< 400). - Consider costs, computer capabilities, time available, and technical assistance available in choosing a model.	- Conduct a unidimensional study (consider the modified parallel analysis method or non-linear factor analysis). - Carry out a classical item analysis to help in model selection(consider sample size also in model selection).
Model Features	- Identify key demographic variables in the population of interest (e.g., race, sex, or geography) and compare item parameter estimates in these subgroups. (Use a t-test, or the Mantel-Haenszel statistic.)	- Identify key demographic variables in the population of interest and conduct item bias investigations (e.g., plot method or total area). - Check the invariance of ability estimates across subsets of items (e.g., hard vs. easy).
Model Predictions	- Look at residuals by ability level (across items) and by items (across ability levels).	- Compare residuals and standardized residuals for several models for both items and ability levels. - Check a variety of hypotheses about the fit (e.g., look for context effects, correlates with fit statistics such as item format, and item content). - Determine the practical consequences of misfit on the intended application (usually through simulation techniques).