Calibration Efficacy of Three Logistic Models to the Degrees of Reading Power Test Using Residual Analysis

by

Monique V. Granville

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
in
Educational Research and Evaluation

Approved

_____	_____
Lawrence H. Cross, Chairman	Alexander Bruce
_____	_____
Robert J. Harvey	Robert S. Schulman
_____	
Kusum Singh	

November 18, 1997

Blacksburg, Virginia

Calibration Efficacy of Three Logistic Models to the Degrees of Reading Power Test Using Residual Analysis

by

Monique V. Granville

Lawrence H. Cross, Chairman

Keywords: Degrees of Reading Power (DRP), Literacy Passport Test (LPT), residual analysis, Item Response Theory (IRT), goodness of fit

(Abstract)

The publisher of the Degrees of Reading Power test of reading comprehension (DRP) calibrate their test using an item response model called the Rasch or one-parameter logistic model. The relationship between the use of the Rasch model in calibration of the DRP and the use of the DRP as a component of the Virginia Literacy Passport Testing Program (LPT) is addressed. Analyses concentrate on sixth grade students who were administered the DRP in 1991.

The question that arises is whether the Rasch model is the appropriate model to use to calibrate the DRP in this high-stakes setting. The majority of research that has been reported by the publisher of the DRP to assess the adequacy of the Rasch model have not included direct checks on model assumptions, model features or model predictions. Instead, they have relied almost exclusively on statistical tests in assessment of model fit. This study will assess the adequacy of fitting DRP test data to the Rasch model through direct examination of the assumptions, features and predictions of the IRT model. This is accomplished by comparing the Rasch model to the less restrictive two- and three-parameter logistic models. Robust IRT-based goodness-of-fit techniques are conducted.

When the DRP is used in a high stakes setting, guessing is likely for those in jeopardy of failing. Under these circumstances, we must attend to the possibility that guessing may be a factor and thereby calibrate the DRP with the three-parameter model, as this model takes guessing into account.

Table Of Contents

List of Illustrations

Page

Page

List of Tables

Chapter I

Introduction

The publisher of the Degrees of Reading Power test of reading comprehension (DRP) calibrate their test using an item response model called the Rasch or one-parameter logistic model. The relationship between the use of the Rasch model in calibration of the DRP and the use of the DRP as a component of the Virginia Literacy Passport Testing Program (LPT) is addressed. Analyses concentrate on sixth grade students who were administered the DRP in 1991.

The question addressed by this research is whether the Rasch model is the appropriate model to use to calibrate the DRP in this high-stakes setting. The majority of research that has been reported by the publisher of the DRP to assess the adequacy of the Rasch model have not included direct checks on model assumptions, model features or model predictions. Instead, they have relied almost exclusively on statistical tests in assessment of model fit. This study assesses the adequacy of fitting DRP test data to the Rasch model through direct examination of the assumptions, features and predictions of the IRT model. This is accomplished by comparing the Rasch model to the less restrictive two- and three-parameter logistic models. Robust IRT-based goodness-of-fit techniques are conducted.

An extension of the problem concerns whether it is appropriate to use the Rasch calibrated DRP as a component of the LPT. The LPT consists of three tests. These tests measure the competency of students in reading, writing and mathematics. Students must pass all three components of the LPT before they can be classified as a ninth grader. Any student who has not passed the LPT by grade eight is classified as ungraded, cannot participate in school sponsored extracurricular activities, and is not eligible for graduation from high school. Students are afforded opportunities to take the LPT until receiving a passing grade in each section.

In order to calibrate a test with the Rasch model, one must be able to assume that the examinees appropriately consider each item and that guessing is minimal. The use of the DRP in a high stakes environment may result in the violation of the assumption of minimal guessing. Because the multiple-choice format is used in the DRP and because multiple-choice tests encourage guessing by having the correct response option available, the assumption of minimal guessing may be at risk. Moreover, since students know that they must pass this exam in order to be eligible for promotion to the next grade, it is likely that guessing will occur, especially among those students in jeopardy of failing.

It is important to determine whether the Rasch model is the appropriate model to use to calibrate the DRP. We must attend to the possibility that guessing may be a factor when the DRP is used in a high stakes settings and thereby calibrate the DRP with the three-parameter model, as this model takes guessing into account.

Chapter II

Review of Related Research

The Degrees of Reading Power

Following the release of the 1984 results of the National Assessment of Educational Progress (NAEP), newspapers all over the country were reporting American school children were seriously deficient in reading, writing and mathematics. Virginia Legislators, concerned about the presumed low level of academic achievement, decided that an assessment of literacy was necessary. In 1986, the Governor's Commission on Excellence in Education in Virginia made recommendations for Literacy Passport Tests in reading, writing and mathematics. The Virginia Department of Education, in turn, mandated standards of learning which defined the foundation for the development of the curricula. A new policy was adopted in 1989 which specified a) the tests provide the basis for determining whether a student would be promoted to the ninth grade and b) that eligibility to participate in extra curricular activities would depend upon passing these tests.

The reading section of the LPT consists of the DRP test. There are three levels of the DRP. The Primary and Standard DRP tests measure reading comprehension whereas the Advanced tests measure higher order processing such as analyzing, deriving and/or combining text propositions. Only the Standard form is used in the Virginia LPT program. The DRP is a multiple-choice (modified cloze) test of reading comprehension which is standardized and norm-

referenced. It is designed to provide a measure of a student's ability to understand the messages contained in English text.

An unbiased and interpretable reading comprehension test is one in which the test passages provide the content knowledge necessary to respond to the question, the difficulty of the question is linked to the difficulty of the prose, the questions require that the prose must be read, there is clearly only one correct answer, and the task should disrupt the reading process as little as possible. To meet these requirements, it is necessary to write prose directly instead of using existing material. These conditions provided the basis for the development of DRP prose (Koslin, Zeno, Koslin, Wainer and Ivens, 1987).

The DRP is similar in construction to the Cloze method of measuring reading comprehension. Cloze items introduced by Taylor (1953) are formed by deleting every nth word or randomly deleting a fixed proportion of words. When a reader encounters a blank space, he or she must think of possible words that might fit in the blank. The use of the surrounding context helps the reader guess the missing word. The DRP however, consists of a series of passages that have essential words deleted rather than words deleted at random. When a student encounters a sentence where a word has been deleted, he/she must choose from among five options to indicate which word best fits in the blank. All options are semantically plausible and syntactically correct. The DRP is designed so that the entire passage must be understood before the examinee can choose the correct response. The material focuses on a variety of subjects presumed to be relatively unfamiliar so that the selection of the correct response will not be a function of prior knowledge.

Prose difficulty can be indexed by the number of surface features descriptive of difficulty. The relationship between these surface features and prose difficulty have led to the development of readability formulas used to

predict the difficulty of reading materials and text. The organization of the DRP is such that the readability of the test passages is ordered from high to low. This design feature of the DRP is optimal in the application of readability formulas because these formulas index text difficulty through the assumption of well-developed, organized and coherent prose. For the DRP, readability is measured in terms of DRP units. This scale ranges from 0 to 100 units where high DRP units reflect more difficult prose or low readability. Readability is calculated based on a transformation of the Bormuth (1969) formula. This formula is also used to provide a readability index measured in DRP units for educationally relevant material.

A problem with traditional multiple-choice tests of reading comprehension is that it is impossible to precisely link difficulty of the multiple-choice question to the difficulty of the passages on which they are based. Easy questions can be asked about a difficult subject and hard questions can be asked about a relatively simple subject (Bormuth, 1970). Because multiple-choice questions sometimes provide clues to the answer, they can be answered correctly at a better than chance level and without reading the text (Tuinman, 1974; Powers & Leung, 1995). Authorities agree, that if students are able to answer a question without reading the passage, then the test is not measuring comprehension (Schell, 1984). These criticisms are largely overcome by the DRP because in addition to linking the difficulty of the question to the difficulty of the prose, Cloze items require reading in order to respond correctly. There is a general acceptance in the reading literature that Cloze items measure reading comprehension.

Applicability of the Rasch Model to the DRP. The Rasch model assumes that a) all items have the same relationship (equal discrimination indices) to the underlying latent trait (reading ability) and b) the probability that examinees of

very low ability will get an item right is zero (minimal guessing). Both assumptions are almost always unrealistic with multiple-choice tests.

The notion of equal discrimination indices is so restrictive that this assumption is viable only when items have specifically been chosen to meet this criteria, (Birnbaum, 1968; Hambleton & Traub, 1973; Lord, 1968) and is questionable for many types of tests (Wood ,1978; Gustaffson, 1980). Empirical evidence has so often shown that acceptable achievement items correlate with varying degrees to the underlying trait that a parameter describing item discrimination is necessary and sufficient and that these indices will be unequal (Traub, 1983).

Proponents of the Rasch model believe that the solution to differing item discriminations is to exclude 'nonfitting' items (Ryan, 1982). Several researchers have commented on the notion of fitting a model through the elimination of items that do not fit the model as being untenable (Traub, 1983; Gustafsson, 1980). Gustafsson (1980) found that item elimination for the purpose of model fit results in the proportion of misfitting items to be too large to provide satisfactory results. Likewise, Lumsden (1980) found deleting data to obtain fit will result in the possibility of the obliteration of a systematic subsample.

It is widely recognized that for multiple-choice tests, when examinees encounter items for which they don't know the answer, they are likely to guess. Because passages of the DRP are presented in increasing order of reading difficulty, one may reasonably assume that when an examinee encounters passages beyond their reading level, the potential for guessing increases. Advocates of the Rasch model conclude if the difficulty of a test is matched to the ability of the examinees being tested then the examinees will have no need to guess and the assumption of minimal guessing can be supported. The foregoing argument seems to suggest that the Rasch model can only be used when

multiple-choice items are easy for the examinee population. This suggestion infers that the Rasch model will only provide invariant item parameter and ability estimates when examinee populations do not range widely in ability level (Wright, 1968). Cross (1995) attributes large increases in DRP scores reported for normative data from New York between the 1981-82 and 1987-1988 norming studies to guessing. Between the above mentioned norming studies, there was a switch in the use of the DRP as an assessment tool to that of a high stakes test.

Many studies have found the Rasch model inappropriate for multiple-choice tests. Andersen (1973) found the Rasch model did not fit the verbal part of the Scholastic Aptitude Test because of unequal discrimination indices. Hambleton and Murray (1983) found the three parameter logistic model to fit NAEP mathematics data better than the Rasch model. Divgi (1986) found the Rasch model inappropriate for seven popular multiple-choice reading tests; the California Achievement Tests; Comprehensive Test of Basic Skills, Iowa Test of Basic Skills, Metropolitan Achievement Tests, Sequential Tests of Educational Progress, Stanford Achievement Tests and Gates-MacGintie Reading Tests. He concluded that the Rasch model should never be used for multiple- choice tests.

These considerations suggest that more than one parameter may be necessary to calibrate multiple-choice achievement test scores. A discrimination parameter is needed to account for varying correlations between the latent trait and item responses, a difficulty parameter is required to index the difficulty of an item and a parameter that accounts for the effects due to guessing must be introduced.

The publishers of the DRP report that the California Achievement Test (CAT) correlates .85 with the DRP. In order to confirm the concurrent validity of the DRP, Miller (1988) correlated the esteemed Stanford Achievement Test

(SAT) and the DRP for four samples of middle school students who took both tests. The correlation between the DRP and the SAT scores failed to approach the level achieved by the CAT but ranged between .72 to .76. These DRP estimates of validity as reported by the publisher as well as by Miller (1988) are impressive. However, as many school districts are using the DRP to group students into academic classes, Miller (1988) wanted to determine how well the DRP actually predicts student achievement. He found both instruments provided poor to moderate correlations (.20 to .68) with final first-semester grades in science, language arts and social studies. The highest correlations tended to be obtained for science grades and the lowest for grades in social studies.

By comparing students' DRP unit test score representing the reader's independent reading level with the readability values of books that students had actually read, Snyder (1993) found students who performed well on the DRP capable of reading books within the level as predicted by the DRP readability index. However, students scoring low on the DRP were capable of reading and comprehending material far beyond that as predicted by the DRP. Carver (1985) reports that the DRP tends to underestimate the ability of readers in the lower grades and overestimate the ability in the upper grades. Bormuth (1985) argues that although reading achievement tests are designed to avoid measuring background knowledge, the omission in his formula to account for background knowledge may cause this bias. Bormuth explains that in grades 1 through 6, a reader's knowledge is far greater than one's reading ability. The Bormuth scale is not capable of gauging background knowledge so student scores are underestimated. On the other hand, in grades 8 through 12, students' reading ability is near mastery while their knowledge is scattered so we would expect students to be much less able to understand material than their reading scores

would predict. Another possible explanation is that the Rasch model is not providing adequate item parameter and ability estimates because the model cannot adequately handle guessing or shifting discriminations.

Item Response Theory

The IRT models used with dichotomously scored multiple-choice test data are the one-, two- and three-parameter logistic models. All three models use item characteristic curves (ICCs) to relate the probability of responding correctly to an item, $P_i(\theta_j)$, to the ability of the examinee, θ_j. This probability is a monotonically increasing logistic function of $(\theta_j - b_i)$, where b_i is the difficulty of item i.

Item Response Models

The one-parameter model (named the Rasch model in honor of its inventor) (Rasch, 1960, 1966) assumes that the probability that an examinee will correctly answer an item is a function only of the ability level of the examinee and the difficulty of the item. More specifically:

$$P_i(\theta_j) = \frac{1}{1 + \exp -(\theta_j - b_i)} \qquad (1)$$

where i = 1,..., k items; j = 1,..., n examinees.

Figure 1 provides examples of three typical one-parameter model ICCs. These ICCs consist of nonintersecting curves and differ only in their position (location) along the ability scale. Difficult items are located to the right and easy

items are located to the left on the ability scale. This model specifies that the probability of correctly responding to an item is a function a single parameter, the difficulty of an item. The difficulty of an item, as defined in IRT, is the center

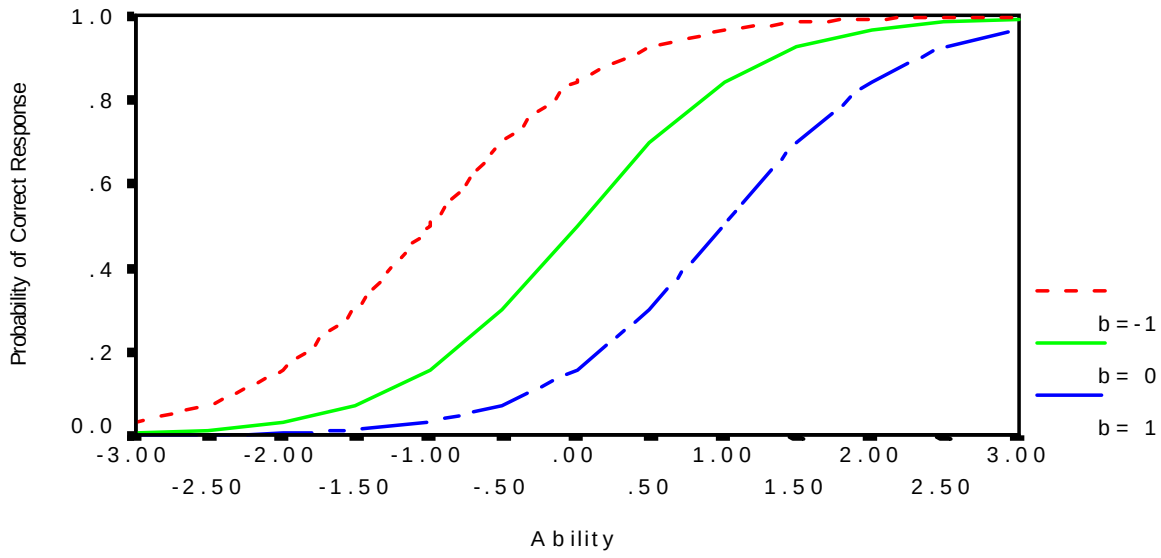

Figure 1. One-parameter ICCS for three test items.

of the item response curve. If we mathematically find this center then the difficulty is defined as the ability level associated with a probability of .50. The value of this parameter determines the location of the ICC in relation to the ability scale. As items get harder, the ICC shifts to the right and examinees require increasing levels of ability in order to have a 50% chance of getting the correct response. Note that the lower asymptote of the ICC is zero. This means that examinees of low ability will have a zero chance of correctly responding to difficult items. Thus, no allowance is made for the possibility that low scoring examinees may guess.

The two-parameter model (Lord, 1952) includes a discrimination parameter (a_i) which indexes the strength of the relation of the item response to the latent trait. The two-parameter model may be expressed as follows:

$$P(\theta_j) = \frac{1}{1 + \exp - [a_i(\theta_j - b_i)]} \qquad (2)$$

where i = 1,...,k items; j =1,..., n examinees.

Figure 2 provides illustrations of three ICCs for the two-parameter logistic model. The ICCs of the two-parameter model have slopes that increase at different rates and vary in

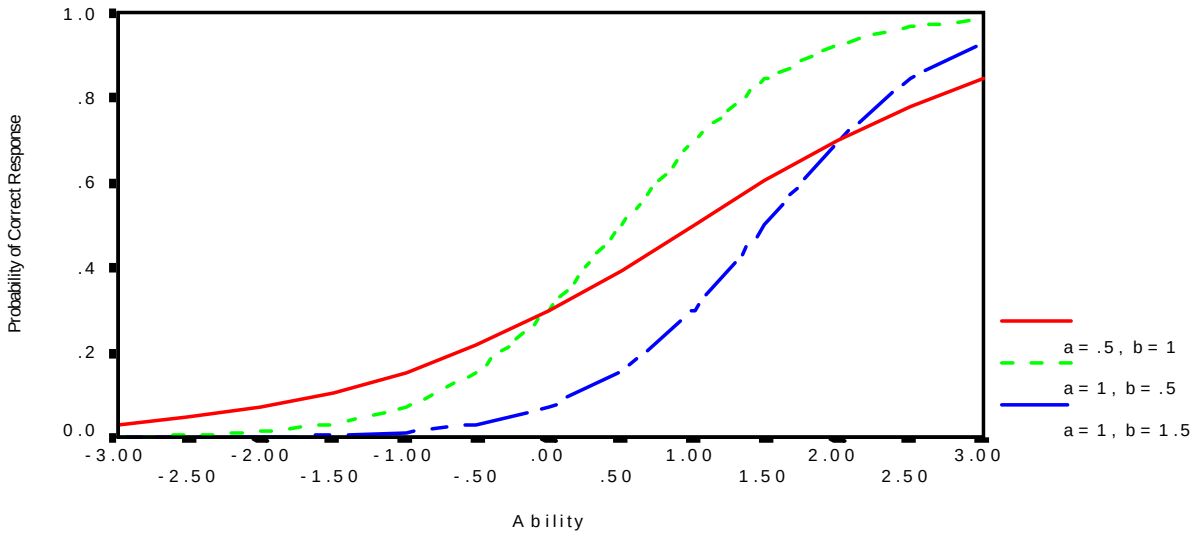

Figure 2. Two-parameter ICCS for three items.

translation along the ability scale. Intersecting curves denote that some items are less discriminating than other items.

The center of the ICC is the point at which the acceleration of the function changes from a positive to a negative acceleration. It is at this point that the slope of the curve is at a maximum. The a parameter is proportional to this

maximum slope at the point b on that ability curve. High values of the a parameter result in steeper ICCS. Items with steeper slopes are more useful in distinguishing examinees by ability (within the range of ability where the slope is steepest) than items with less steep slopes. The larger the item discrimination parameters, the stronger the dependence of $P_i(\theta_j)$ on $(\theta_j - b_i)$. Thus, within the range of ability, large item discrimination coefficients indicate that the performance on these items more closely gauges examinee ability than items with low discriminations. Whereas theoretically the item discrimination parameters are without bound, practically these values range between zero and two. For all ICCs in Figure 2, the probability of correctly responding to an item is .5 when $\theta = b_i$. It follows that the b parameter is the item difficulty parameter as well as the point on the ability scale at which examinees have a 50 percent chance of a correct keyed response.

If there is the possibility that an examinee may correctly respond to an item through guessing, a third parameter may be introduced into the two-parameter logistic model. This three-parameter logistic model may be expressed as follows:

$$P_i(\theta_j) = c_i \frac{1 - c_i}{1 + \exp - \left[a_i(\theta_j - b_i)\right]} \qquad (3)$$

where i = 1,...,k items; j =1,..., n examinees.

The three-parameter model (Birnbaum, 1968) is indexed by item difficulty, item discrimination and a lower asymptote or guessing parameter, c_i. The c parameter is the height of the lower asymptote of the ICC and is introduced into the model to account for the performance of low ability examinees on multiple-choice items. The value of the lower asymptote is typically lower than the probability of correctly responding to an item if an examinee randomly guessed.

Figure 3 shows three test items that differ in difficulty, discrimination and lower asymptote. These curves intersect, vary in translation along the ability curve and have lower asymptotes considerably above zero. The b parameter for a test item is the point on the ability scale where the slope of the ICC is a maximum. For the three-parameter model, the slope of the curve at b is equal to $.425a_i(1-c_i)$. The probability of a correct response at b on the ability scale is (1 + c)/2. When c = 0, the probability of a correct response at b is 50 percent, when c is greater than zero, this probability exceeds 50 percent. The a parameter is proportional to the slope of the ICC at the point b on the ability scale. The steeper the slope, the higher the a parameter.

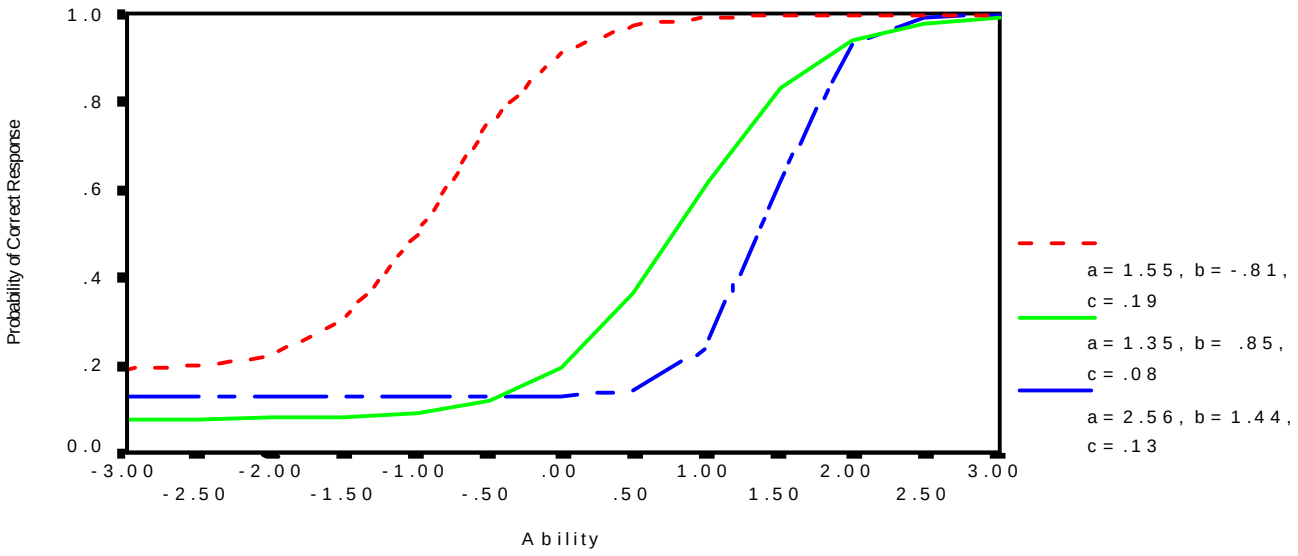

Figure 3. Three-parameter ICCS for three items.

Although the three-parameter model accounts for guessing, this model has a limitation in the manner in which this is accomplished, (Birnbaum, 1968). The model is interpreted as being one which assumes that examinees either

guess randomly among all choices or answer correctly on the basis of their knowledge. This interpretation leaves out the possibility of guessing correctly to an item after being able to eliminate one or more alternatives, (partial knowledge guessing).

Model Assumptions

In order to use the one-, two- and three- parameter logistic models to calibrate test scores, certain assumptions must be made about the test, the data and the model. All three models require the test to be administered under nonspeeded conditions and the test to be unidimensional. For the one- and two-parameter models, we must assume that guessing is minimal and for the one- parameter model we must additionally assume that the discrimination indices are equal.

Unidimensionality. Unidimensionality implies that the performance of an examinee is attributable to one and only one underlying trait. Local independence means that the probability of an examinee correctly responding to an item is unaffected by responses to other items on the test. Statistically, this means that the items are uncorrelated for examinees at the same ability level. For a fixed ability group, if the items were not statistically independent, then some examinees would differ in their expected test scores. More than one ability is then necessary to account for differing test performances. Therefore, a necessary and sufficient condition for the existence of local independence is a unidimensional latent space.

Nonspeeded Test Administration. When a test is timed, some students may not have sufficient time to complete the exam in which case the test measures response speed as well as the latent ability being measured. To the extent that more than one trait underlies test performance, the unidimensionality assumption is violated. Thus, an implicit assumption of the unidimensional IRT model is that the test must be administered under nonspeeded conditions.

Equal Discrimination Indices. Only the Rasch model requires one to assume equal discrimination indices. Discrimination indices measure the extent to which a certain item is able to discriminate between examinees with varying degrees of knowledge. For the Rasch model, this index is one for every item indicating that the only thing that matters when responding to an item is the ability of the examinee and the difficulty of the item.

Minimal Guessing. The assumption of minimal guessing is most plausible with free response item types, when the test has a difficulty level lower than the difficulty level of the examinee with the lowest ability, or in situations where a test is administered to students following effective instruction. The guessing problem arises when examinees are not given explicit directions for what to do when he or she does not know or is not sure of the correct answer. The one- and two-parameter models do not contain a parameter that takes guessing into account. These models, therefore, assume minimal guessing.

Advantages of IRT Models

Interest in IRT stems from three desirable features that are obtained when a model fits a test data set:

1. An examinee's estimated ability does not depend on the particular sample of test items chosen from the total pool of items.

2. Item statistics (e.g. item difficulty and discrimination indices) are not dependent upon the particular sample of examinees drawn from the population of examinees from whom the test was intended.

3. A statistic indicating the precision of the estimated ability is provided for each examinee. This estimate is a function of examinee ability and the statistical properties of the items in the test.

A poorly fitting model will not yield invariant item and ability parameters. In any application of IRT, it is important to determine whether invariance holds since this property is the cornerstone of IRT.

Ability and Parameter Estimation

To use the logistic models, it is necessary to obtain estimates of the ability of each examinee and the parameters associated with each item. Computer programs have been developed to obtain these estimates iteratively. The parameter estimation process results in test items and examinees being placed on an ability scale in such a way that the difference between expected and observed probabilities of correctly responding to an item is small. To increase

precision, item parameter and ability estimates are revised until maximum agreement is obtained between parameter estimates and the actual test data.

In general, the programs provide accurate estimates for large sample sizes and long tests. Standard errors of the predicted estimates, item parameter estimates for all items and ability estimates for every examinee are provided in program output. A goodness-of-fit test is available with some programs. The goodness-of-fit statistic determines whether the estimated ICC (based on observed proportions correct by ability level) for each item conforms to the predicted ICCs (based on estimates of the parameters) in Equations 1, 2, and 3.

Item Response Theory vs. Classical Test Theory

Classical test theory (CTT) is a model that represents the way in which errors of measurement influence observed scores and the quantitative aspects of test efficacy such as reliability and validity. With the classical test model, an examinee's true score is defined as his or her expected test score over repeated administrations of the same (or parallel) test. It can never be measured directly and therefore the quantity is deemed latent. The observable quantity that is measured is the total score or total number of correct responses on the particular test. CTT can be applied to tests which are assumed to have one or more traits that underlie test performance. There are seven basic assumptions of CTT (Allen and Yen, 1979):

1. CTT is based on an additive model in which an examinee's observed score is composed of a true score plus random error.

2. Assuming average random error is zero and there are no practice or fatigue effects, if an examinee were to take a test an infinite number of times, his or her mean observed score would equal his or her mean true score.

3. The correlation between true and error scores is zero.

4. The correlation between error scores on parallel forms is zero.

5. Parallel test forms have equivalent content, equal observed score means, variances and reliabilities.

6. Test forms are true score equivalent if their true scores are the same except for an additive constant.

7. If an observed score is used to predict a score on some criterion measure, the correlation of error scores with the criterion measure is assumed to be zero.

In CTT, the statistical characteristics of the test depend upon the statistical characteristics of the items used to build it and the examinees who take it. These characteristics vary according to the level and heterogeneity of ability among the group tested, and test length, all of which may vary from one occasion to the next. This lack of invariance is one of the leading criticisms of CTT.

The problem with the item parameters in CTT (e.g. item difficulty, item discrimination and reliability) is that they depend on the sample from which they were drawn. For example, if a test was administered to a group of high ability students, item difficulties are very different from those obtained if the test is administered to a group of examinees of low ability. Item difficulty will be high when administered to a high ability sample and low if the item is administered to a sample of low ability students. Because the item discrimination index is a correlation and correlations are effected by the heterogeneity of the sample, item discrimination indices will tend to be higher for samples that are heterogeneous in ability. As a result, item statistics are only useful for item selection when tests are administered to samples similar to the sample of examinees from which the statistics were drawn. The ability to generalize results to other populations is limited.

Another shortcoming of CTT concerns the difficulty to develop parallel forms resulting in the underestimation of alternate parallel forms reliability. In CTT, typically it is assumed that the standard error of measurement is the same for all examinees. The problem here is that we know that some people perform tasks more consistently than others and that high ability people are more consistent than low ability people (Hambleton and Swaminathan, 1985). In addition, test scores are dependent on the difficulties of the items on a test, but one takes item difficulty into account when building parallel forms. The easier the test, the higher the p-values, the higher the total score. This may effect the

criterion related validity of the test. However, test score validity can be increased if test difficulty is matched to the approximate ability of the examinee.

For years, measurement specialists have recognized the shortcomings which manifest themselves as technical problems in the design of tests, identification of biased items and equating test scores. Due to these constraints, measurement specialists have developed new test theories which overcome these limitations. One theory that has commanded considerable interest is item response theory.

An item response model assumes that there is a relationship between observable responses to test items and a latent trait which is assumed to underlie test performance (Lord and Novick, 1968). The relationship between the observed performance and the latent trait is described by a probabilistic mathematical function called an item characteristic function.

IRT has a number of advantages over CTT approaches. IRT item parameters (discrimination, difficulty and pseudo-guessing) are invariant across samples. This means the ICC will have the same form in any subpopulation if the probability of correctly responding to an item is a function of a single latent trait. Theoretically, the parameters of IRT are sample invariant within a scaling factor. Therefore, parameters can be estimated from one sample and with the use of a scaling constant can be applied to other samples. This implies that they are not dependent on the ability level of the group on which the item parameters were developed. IRT models also produce invariant ability estimates. Item

parameters can be estimated by having a sample of examinees answer a sample of items.

Another distinction between IRT and CTT is the concept of standard error of measurement (SEM). A major advantage with IRT is that the SEM differs for different individuals. The smallest SEM is observed at the point where the test provides the maximum amount of information.

A practical benefit of the IRT model is that it is possible to determine the probability that an examinee will correctly respond to a test item. This information is valuable to test developers who are attempting to design tests for specific populations. IRT has also provided better solutions to other testing problems such as designing tests, the identification of biased items and the equating of test scores. The design of tests is facilitated as IRT models can provide the location of the maximum discriminating power of items. In addition, because IRT models can adequately handle the differences in ability among the examinees, the identification of biased items and the equating of test scores is less difficult than are attempts using classical methods.

In summary, the advantages of using IRT over CTT stems from three desirable characteristics of IRT. Scores describing examinee proficiency are not sample dependent, item statistics are not sample dependent and IRT models provide a basis for matching item difficulty to ability level.

Assessment of Model-Data Fit

Hambleton and Rogers (1986) have outlined three approaches for conducting effective goodness-of-fit studies: (1) checking model assumptions (unidimensionality, equal discrimination indices, minimal guessing and nonspeeded test administration); (2) checking model features (investigations into the attainment of invariance properties); and (3) checking model predictions (model-test data fit studies). Several Studies (Hambleton, 1989; Hambleton & Swaminathan, 1985) have detailed methods to investigate the viability of each of the three types of evidences.

Model Assumptions

Assumptions that must be checked for all three models are whether the test is nonspeeded and whether the data are unidimensional. For the one- and two-parameter models, we must additionally check that guessing is minimal and for the one-parameter model, equal discrimination indices. When the assumptions of a model are not met in the data, the reliability of parameter estimates is questionable. For this reason, it is necessary to confirm the viability of the assumptions of each model of interest.

Unidimensionality. Many researchers believe that when one considers all the skills necessary to solve all of the items on a test it is concluded no single trait can account for the item responses of every examinee (Hambleton & Swaminathan, 1985; Humphreys, 1985,1986; Reckase, 1979; Stout, 1987; Tatsuoka, 1990; Traub, 1983; Yen, 1985) and that all tests are likely to be multidimensional to some degree (Humphreys, 1985; Reckase, Ackerman & Carlson, 1988; Wang, 1988; Yen 1985). The recognition that examinee performance is confounded with many cognitive and affective factors such as fatigue, test wiseness, cognitive style, motivation and anxiety has led to the development of numerous techniques which assess the dimensionality of a test. Although this assumption provides the basis for most mathematical measurement models, there is no consensus about which methodology is best suited to assess the dimensionality of test data (Berger & Knol, 1990; Divgi, 1981; Hambleton & Rovinelli, 1986; Hulin, Drasgow, & Parsons, 1983; Zwick, 1987).

In the application of IRT models, the generally accepted definition of unidimensionality refers to demonstrating the existence of one dominant dimension underlying the responses to a test (Nandakumar, 1994). Hattie (1985) summarized over 30 indices for assessing unidimensionality. Hulin, Drasgow and Parsons (1983) also supply a review of existing methodologies. Among those, the traditional approaches are factor analytic requiring the assessment of dimensionality based on sample eigenvalues calculated from either phi or tetrachoric correlations. Current methodologies point to problems

associated with the use of factor analysis on dichotomously scored items. The problem is that factors related to the difficulty of the items arise when using factors analysis on binary data (Hertzman, 1936; Spearman, 1927). Wherry and Gaylord (1944) contended that difficulty factors result when a matrix of phi correlations are used in the factor analytic procedure and if difficulty factors are also obtained when using tetrachoric correlations then the data set is multidimensional. A problem associated with tetrachoric correlations is that they can be affected by guessing (Carroll, 1945; Lord, 1980) and non-Gramian matrices may result when factor analyzed.

The treatment of IRT as a special case of nonlinear factor analysis is gaining popularity in educational measurement. De Champlain (1995) reports that several authors (Balassiano & Ackerman, 1995a, 1995b) have shown nonlinear factor analytic models to be mathematically equivalent to the logistic functions. It, therefore, makes sense to examine the dimensionality assumption within the context of NLFA . Three commercially available programs exist, however, the weaknesses associated with their use are currently under review and is beyond the scope of this study to attempt to reconcile. Because of this, the traditionally used classical factor analysis of phi coefficients is used to determine dimensionality.

Nonspeeded Test Administration. A test is a pure power (nonspeeded) test if all examinees are given adequate time to respond to all items. The total number of incorrect responses is thereby equal to the number of items to which the

examinee provides an incorrect response. A test is speeded if some examinees do not have adequate time to complete all the items within the allocated time. In a pure speed test, incorrect responses are equal to the number of unattempted items. In order to determine the extent to which a test is speeded, an assessment of the number of examinees who fail to complete all the test items is necessary.

When most of the examinees answer most all of the items, speed is not relevant to test performance. To determine the extent to which a test is speeded the percentage of students who compete the exam, who complete 75% of the exam, and the number of items completed by 80% of the examinees (Swineford, 1956) is calculated.

Equal Discrimination Indices. In the Rasch model, it is assumed that the only characteristic of an item that influences examinee performance is the difficulty of the item. The implication is that there is a negligible difference in discrimination among the items in a test.

The assumption of equal discrimination is viable when the distribution of item-total test score biserial or point-biserial correlations is reasonably homogeneous. Hambleton and Swaminathan (1985) suggest acceptable levels are those falling within $\pm$ .15 of the average item-test score correlation. A large percentage of correlations falling outside this range is an indication of unequal discrimination indices.

If the range of the item biserial correlations is large, uniformity of discrimination does not hold (Hambleton, 1989) and histograms of the discrimination indices estimated for the two- and three- parameter models will show a large range. As cited in Hashway (1978), Kifer, Mattson and Carlid (1975) consider IRT item discrimination indices of plus or minus three standard errors falling within the range of .80 to 1.20 to provide an adequate confidence interval around unity to justify that the assumption of equal discrimination indices has been satisfied.

To determine if the assumption of equal discrimination indices is viable, the range of the item point-biserial correlations is calculated. If this range is large, the two- or three-parameter models provide a better fit to the test data. IRT item discrimination indices falling outside of three standard errors are noted to help identify items which fail to satisfy the criterion of equal discrimination.

Minimal Guessing. Numerous techniques have been proposed to assess the effect of guessing on test performance. In a review of these, Rudner (1983) describes two categories: a) those based on item response patterns and b) those based on item response theory.

Ozcelik and Berberoglu (1991) compared the difference between mean item difficulties adjusted for guessing (p_c) and the observed mean item difficulties (p) to test the assumption of zero chance level. If this difference is approximately zero, for high, middle and low ability groupings, the assumption of minimal guessing is viable.

If low ability students are getting the answers to the hardest questions correct, this is evidence that the assumption of minimal guessing has been violated. As cited in D'Costa (1993), the Sato caution index (CI) (1980) was developed to provide a measure of an examinee's inconsistent response patterns. The Sato index is based on the difficulty of items, the number and type of "errors" committed by an examinee. These errors are defined as items which are answered correctly although the item was not within the subject's ability level and items for which an incorrect response was given although the item was within the subject's ability level. The CI is calculated as the difference between the number and difficulty of items gotten wrong within an examinees' ability level (concern) and number and difficulty of items gotten right that is outside of one's ability (surprise).

Both of these types of responses are considered unusual. When a student misses an item within his/her ability level this may be indicative of carelessness. When a student gets an item correct that is above his/her ability level this may be indicative of guessing. Sato (1980) recommends that a CI greater than .5 represent significant "caution". D'Costa (1993) advocates analyzing the components of the CI separately because combining the indices may not always uncover inconsistent response patterns. Separating these indices revealed that the CI may be below the recommended cutoff of .5 while the separate indices are above the cutoff.

The difference between observed and adjusted difficulties are compared for high, middle and low ability groups. Small differences are an indication that

the assumption of minimal guessing has not been violated. Under the assumption of minimal guessing, a histogram of the guessing parameters obtained from the three-parameter BILOG run will reveal c parameters scattered about zero with small standard errors. An analysis of the Surprise (B) index helps to pinpoint examinees with inconsistent response patterns.

Model Features

A poorly fitting model will not yield invariant item parameter estimates or statistics that accurately describe the items. Previous studies (Hambleton & Murray, 1983; Hambleton & Rogers, 1986; Hambleton, Rogers & Arrasmith, 1986) have employed variations of the plot method (Bejar, 1980) to check the property of invariance. Using some independent variable (for example, ability) the data is dichotomized. The sample of high and low ability examinees is then randomly divided into half again forming two randomly equivalent samples of high ability examinees and two randomly equivalent samples of low ability examinees. For each of the four samples, the one-, two- and three-parameter logistic analyses are conducted. Plots of item statistics from the high and low ability samples serve as a baseline when comparing the shape of the plots to the four subsamples. If the plots are similar, then the groups are randomly equivalent and perform no different. The difference between the subgroups (ability) can then be ruled out as a factor in the performance of those items. If the plots are not similar in shape, this indicates that the independent variable is a possible biasing factor. The next step is to use the item statistics to identify those items which are producing inconsistent performance and further review these items as being the possible cause of bias.

To assess the invariance of item parameter and ability estimates, a modified Bejar (1980) method is used. Items are calibrated on high and low scorers. If there is no effect in parameter estimation due to ability, a plot between item parameter estimates and ability will reveal a linear relationship. Likewise, to assess the viability of the invariance of ability estimates, ability is calculated separately on easy and hard items. If a linear relationship exists between ability estimated for the two samples, this is indication that ability is invariant to the sample of items administered. These relationships are explored for the one-, two- and three parameter logistic models.

Model Predictions

Residual analysis offers a more practical means of examining goodness-of-fit through revealing patterns of misfit (Traub and Wolf, 1981), and offers the most promising approach for addressing model-data fit (Hambleton & Rogers, 1986). Several researchers (Hambleton & Swaminathan, 1985; Hambleton, Murray & Williams, 1983; Kingston & Dorans, 1985, Ludlow, 1986; Murray & Hambleton, 1983; Wright & Stone, 1979) have performed investigations of residuals and standardized residuals to determine model fit to test data. Murray and Hambleton (1983) describe five steps needed to conduct an analysis of residuals:

(1) An IRT model is chosen and parameter and ability estimates are derived from the data.

(2) Ability and parameter estimates are substituted into the logistic model to obtain predictions.

(3) The discrepancies between observed (O_{ij}) and predicted (E_{ij}) performances are examined.

(4) Fit of the model is determined by the size and direction of the residual, ($R_{ij} = O_{ij} - E_{ij}$).

(5) Residuals are plotted as a function of ability to determine specific sources of misfit.

A measure of fit between the estimated ICC and the observed data is done by comparing average item performance for each ability group to the performance level predicted by an estimated ICC. The difference between the observed and predicted proportion correct at a given ability level is called the raw residual.

Ability groups differing in sample size may have the same raw residual for some item. This is an indication that the model data fit is the same in the two ability groups. Suppose two ability groups contain sample sizes 20 and 200. Clearly, the ability group containing 200 examinees will provide for a more accurate estimation of the predicted proportions correct. To provide an assessment of model-fit while taking into consideration differing sample sizes,

Blalock (1979) devised the standardized residual (SR). The SR is computed by dividing the residual by the corresponding standard error of the predicted proportion correct.

To determine model-data fit, raw and standardized residuals are examined in a variety of ways. Large residuals indicate that the chosen IRT model does not fit the data. The relative sizes and distribution of the residuals indicate the degree of model-data fit. For the three models, a comparison of a crosstabulation of item difficulty by the size of the residuals establishs which model has provided the best estimate of performance. For the one-, two and three- parameter models, plots of absolute-values of standardized residuals versus the point biserial correlations reveals whether a model with varying item discriminations is appropriate. Graphing the standardized residuals against ability for individual test items may reveal specific sources of misfit.

Chapter III

Methods

The aim of the present study is to investigate the appropriateness of using the Rasch model to calibrate the DRP test of reading comprehension. The one-parameter Rasch model is compared to the two- and three-parameter logistic models by testing (1) whether the assumptions held by the pertinent model can be justified (2) the extent to which the property of invariant item and ability estimates is achieved and (3) whether the performance predicted by the one-, two- and three-parameter logistic models are consistent with observed responses.

Subjects and Administrative Procedures

The study is based on a random sample of the responses of 2000 sixth grade students residing in the state of Virginia who took the DRP in 1991. The Standard form of the DRP was administered to all sixth grade students. The test was nonspeeded. During 1991, there was no immediate penalty for not passing.

The data were obtained from the Virginia Department of Education on tape. The data appear on the data tape in two different ways: (1) as a vector of dichotomous item scores (1=right /0=wrong) and (2) as a vector of 77 numbers and characters where numbers (1, 2, 3, 4, 5) appear in place of the correct keyed response, letters (a, b, c, d, e,) appear as wrong responses and blanks appear as nonresponses. The keyed responses to the items are identified as another vector of numbers which indicate the answer to each item.

The Degrees of Reading Power Test

The Standard form of the DRP consists of ten passages each containing approximately 250-300 words. In each test booklet, the passages are arranged in increasing order of difficulty. In each passage, seven sentences contain an omitted word. One word is omitted every two to seven sentences. Five choices for each omitted word are supplied in the margin. Each choice is grammatically and syntactically consistent within the sentence. The examinee is required to select the most appropriate word from among the five choices. The test contains a total of 77 items.

Research Questions and Analytic Procedures

The Unidimensionality Assumption.

The assumption of unidimensionality implies that all the items in a given test measure a single underlying trait. Factor analysis is used to assess the dimensionality of the test items. The traditional method is to analyze a matrix of either phi or tetrachoric correlations. Since tetrachoric correlations may not yield a correlation matrix that is positive definite, causing the factor solution to diverge, a matrix of phi correlations is analyzed.

To understand the variance structure among the item scores, one might use either principal components analysis or factor analysis. Since no item represents an error free measure, the choice might be to proceed with a common

factor analysis as the method analyses only the variance common among the items. Because a conclusive test of dimensionality is not available, common factor analysis is compared to principal components analysis in order to determine whether both yield a single dominant factor.

When using component analysis, the total variance is analyzed by inserting unities on the diagonal of the interitem correlation matrix. For common factor analysis, communalities (R^2) are inserted on the diagonal and the derived factors are based only on the variance in an item that is shared with all the other items in the test. To determine the relative importance of the first factor in accounting for the variance associated with the set of variables being analyzed, Reckase (1979) suggests looking for a dominant first factor and a high ratio of the first to the second eigenvalue. A plot of the eigenvalues from largest to smallest is expected to reveal a dominant first factor relative to much smaller subsequent factors. The percentage of variance criterion is also used. No absolute cutting line has been adopted for all data, however, the expectation is to find that the first factor accounts for most of the total or common variance.

Existing methods for assessing unidimensionality are connected to its various definitions in the literature. Reckase (1979) and Drasgow & Parsons (1983) reported that IRT model parameters could be properly estimated with tests deemed unidimensional through less robust definitions. Since other evidences of goodness-of-fit are provided, if the first factor accounts for approximately 20% of the test variance it is taken to mean that acceptable IRT

parameter estimates can be obtained (Reckase, 1979) and the decision is to continue on to other analyses.

The Non-speededness Assumption. When a test is speeded, some examinees will not have enough time to consider items toward the end of the test. To assess speededness, an inquiry into the percentage of "not- reached" items is made.

To analyze speededness item responses is categorized as those which are answered (marked) and those unanswered (left blank) allowing for four mutually exclusive score categories:

R (rights), the number of items marked correctly

W (wrongs), then number of items marked incorrectly

0 (omits), the number of items that are left blank but are followed by items that are not blank, and

NR (not-reached), the number of consecutive items at the end of the test that are left blank.

An important distinction is made between items which have been omitted as opposed to those which have not been reached. Omit implies that the examinee has read the item and decided not to answer the item. Unattempted items are those not reached and therefore are never considered.

To check the extent to which the test is speeded, the approach of the Educational Testing Service (Swineford, 1956) is used. The DRP is judged essentially unspeeded if virtually all of the examinees complete 75% of the test

(reach item 57) and 80% or more of the examinees complete the exam. The DRP is not expected to exhibit signs of speededness.

The Equal Discrimination Assumption. Because the Rasch model has only one parameter, item difficulty, it must be assumed that all the items have the same discrimination index. The assumed value of the discrimination index is one. Only descriptive methods are available to assess the viability of this assumption.

If the test and an item measure the same attributes, performance on the item should be correlated highly with total test scores. If the range of the item-total correlation is large, this is taken as an indication that the items do not have equal discrimination. Histograms of the estimated discrimination indices obtained from the BILOG run for the two- and three- parameter models supplement conclusions based on the range of the point-biserial correlations.

The distribution of point-biserial correlations is observed. If the distribution of point-biserial correlations is reasonably homogeneous then the assumption of equal discrimination holds. If a large percentage of correlations between items and test scores are found to fall inside $\pm$ .15 of the average biserial correlation, the assumption of equal discrimination is considered to hold (Hambleton & Swaminathan, 1985).

To determine misfitting items, an estimate of the logistic discrimination index and its standard error for each item is obtained from BILOG for the two-parameter model. If the estimated item discrimination plus or minus three standard errors is within the range of .80 to 1.20, the item is considered to have

satisfied the item discrimination criterion (Kifer, Mattson and Carlid, 1975). If a large percentage of items fall outside these limits, this means that a model which assumes varying discrimination indices is appropriate.

A large point-biserial range provides an indicating that the fit of the data to the one-parameter logistic model is unlikely. Histograms of the logistic discrimination indices obtained from the two- and three-parameter BILOG calibrations are expected to support this view.

The Assumption of Minimal Guessing. Whereas the three-parameter logistic model accounts for guessing by including a pseudo-guessing parameter, the one- and two- parameter logistic models do not account for guessing and thus their use assumes minimal guessing. There is no direct way to determine if an examinee guesses to an item on a test. However, three indirect methods are employed to assess the extent to which guessing behavior effects test scores on the DRP.

If it can be shown that there is a negligible difference between item difficulties adjusted for guessing and the observed difficulties then the possibility of guessing can be dismissed. The observed item difficulty is the classical p value. The first approach involves the comparison of observed item difficulties (p) with item difficulties that have been corrected for guessing using the correctional formula:

$$p_c = p - \frac{p_w}{a-1}$$

where p_c = the difficulty index corrected for random guessing

p = the uncorrected p value

p_W = the proportion of examinees who attempted the item and missed it

a = the number of alternatives in the item.

High, middle and low ability groups are formed based on BILOG ability estimates from the one-parameter analysis. The uncorrected p-value, adjusted p-value and their differences are tabulated. It is anticipated that the greatest differences between observed and adjusted p values occurs with the combination of the more difficult test items and the less able group. This finding will provide evidence to support the possibility of a violation of the assumption of minimal guessing.

The second approach measures the extent to which examinees are able to correctly respond to items above their ability level. The B index (D'Costa, 1983) is calculated as follows:

$$B = \frac{\text{Sum of q's for items examinee got right beyond the examinee's ability level}}{\text{sum of q's for all items beyond the examinee's ability level}}$$

where q is the proportion of students who responded incorrectly to an item. The B index is computed using a FORTRAN 77 program written by D'Costa (1983). If students with inconsistent response patterns, as evidenced by high B indices, are also low ability students this is considered evidence to further the notion that guessing is a possibility for low scoring students. Comparison of the B index for students arranged by total score is expected to reveal examinees in the lowest ability groups had the most inconsistent response patterns.

Tabulation of the guessing parameters obtained form the three-parameter BILOG run supplements the appraisal of minimal guessing. If guessing is not prevalent, the estimated c parameters will center around zero producing a leptokurtic distribution with a small standard deviation. It is anticipated that the histogram of the c parameters will reveal a large range suggesting that the assumption of minimal guessing has been violated.

Invariance

The major benefit obtained when an item response model fits a test data set is that item parameter estimates do not depend upon the sample of examinees for whom the test is designed and an examinee's estimated ability does not depend upon the sample of test items chosen from the item pool. The attainment of invariant item and examinee ability parameters is evaluated across samples for the one, two and three parameter models.

Item Parameter Invariance. The notion of invariance implies that if an ICC is based on the Rasch model with parameter b_i for some population, then the ICC of any subpopulation calibrated using the Rasch model is indexed by the same parameter, b_i. To determine if the property of item parameter invariance has been obtained, item parameter estimates are compared for ability subgroups for the one-, two and three- parameter models.

Ability groups are formed by dividing examinees into high and low ability groups based on a median split of the number-right scores. The high and low

ability samples are divided again at random into two random halves yielding four groups of examinees (N=500) labeled "High 1", "High 2", "Low 1" and "low 2". BILOG (Mislevy and Bock, 1990) is used to conduct one-, two- and three-parameter analyses for each ability group. The b values estimated within the two high ability groups, and within the two low ability groups are plotted and serve as a baseline against which to judge the adequacy of the invariance obtained when the item parameters are plotted across ability groups. If the invariance of b values holds, similar plots should be obtained. If plots of b values are linear with slopes close to one, then a sample invariant item calibration was approximated.

To examine the relationship between b values, correlations are also obtained between the b values across and within ability groupings. High correlations are an indication that b values obtained in each sample are approximately that same, supporting the assumption of item parameter invariance. Using Fisher's z transformation, a test for the difference between two independent correlations is expected to reveal that the degree of relationship between the b values obtained from the three parameter analysis is significantly higher than those obtained from the one parameter analysis.

Discrimination parameters obtained from the two- and three- parameter analyses are correlated for the two ability groups to assess the viability of the invariance of these parameters. For the three-parameter model, the pseudo-guessing parameters obtained from the two ability groups are correlated to

determine if invariance holds. Since low ability examinees are more likely to guess, the correlation across ability groups is not expected to be high.

Ability Estimate Invariance. Ability estimate invariance implies that no matter which sample of items an examinee responds to, the estimate of that examinee's ability will always be the same within a linear transformation. To investigate the invariance of examinee ability, a comparison is made between ability estimates calibrated from the easiest and hardest items. In this case, the 77 test items are divided into the easiest 38 and hardest 39 based upon a median split of the p values. For the entire sample (N=2000), BILOG ability estimates are obtained separately for each half of the test for the one-, two- and three- parameter models. If plots comparing the ability on hard items versus the ability on easy items are linear with slopes close to one, item invariant measurement of examinees has been approximated. A correlation between these estimates of ability will reveal the extent of the errors in ability estimation. Using Fisher's z transformation, a test for the difference between two independent correlations is expected to reveal that the degree of relationship between the ability estimates obtained from the three parameter analysis is significantly higher than those obtained from the one parameter analysis.

Residual Analysis

Residual analysis offers a means of examining model-data fit through the comparison of observed (actual) performance to the performance expected

under some model. The difference between actual and expected item performance is called a residual. Small residuals are an indication that the model is estimating the performance level of the examinees close to the actual performance of the examinees.

Performance modeled under the one-, two- and three-parameter logistic distributions is compared to the actual performance of the examinees in order to assess which model best fits the data. A scatterplot of the classical item discrimination indices and the standardized residuals will determine if the assumption of equal discrimination indices is appropriate. The association between the percentage of small standardized residuals and classical item difficulty reveals whether the guessing parameter is useful. Items with large (positive or negative) standardized residuals are examined to determine if there are certain characteristics that would cause them to be identified as misfitting.

To calculate the expected proportion of examinees who correctly respond to an item for the one-, two- and three- parameter logistic models, three computer programs written in SAS for the IBM system operating under VM/CMS were prepared. The programs were written to handle dichotomous (right/wrong) data.

Each examinee is placed in an ability group based on the BILOG ability estimate. Ability groups are formed by dividing the ability continuum into 12 equally spaced intervals between -3.0 and 3.0. Examinees with ability estimates falling outside of the minimum or maximum ability continuum are removed from

the analysis. It is expected that this will result in the removal of only a few cases.

Using item parameter estimates obtained from BILOG, the expected performance $\left(\hat{P}_{ij}\right)$ is calculated for the one-, two- and three parameter logistic models. For each model, analyses are based on the performance of ability groups. The midpoint of each ability group will serve as an estimate of the ability of that group.

For each of the 12 ability categories, the average observed performance $\left(P_{ij}\right)$ is calculated as the proportion of examinees in ability category j who correctly responded to item i. The value of P_{ij} is compared to the average probability of answering the item correctly as given by the theoretical ICC. The extent to which the expected performance accurately describes the observed performance is obtained by calculating the difference between the observed and expected proportions forming the raw residual, $R_{ij} = P_{ij} - \hat{P}_{ij}$. There are 924 (77x12) raw residuals computed for each analysis.

The matrix of residuals is summarized in two ways. When the ability group is of interest, the residuals is placed into a 77 X12 matrix based on a subject's ability grouping. The first column, for example, consists of residuals that represent the average performance on each of the 77 items for all subjects in the first or lowest ability group. When generalizations are to be made about specific items or the test, the matrix is transposed. The columns will consist of

the 12 residuals. Each residual is associated with the average performance of each ability group with respect to the particular item.

Standardized residuals (Blalock, 1979) are used to assess model fit, while taking the sample size of the ability group into consideration. The standardized residual is calculated as follows:

$$SR_{IJ} = \frac{P_{IJ} - \hat{P}_{IJ}}{\sqrt{\frac{\hat{P}_{IJ} - \left(I - \hat{P}_{IJ}\right)}{N_{IJ}}}} \qquad (4)$$

where i = 1,..., k items; j = 1,..., n examinees and N_j is the number of examinees in ability category j.

For the one-, two-, and three-parameter models, tabulations of the size and direction of the raw and standardized residuals are compared. Tabulations of the raw and standardized residuals across ability levels for each item will establish for which ability groups the model has problems in estimating performance. Tabulations of the size and direction of raw and standardized residuals across items at each ability level provide insight into which items can be classified as misfitting. A crosstabulation of raw and standardized residuals, across both ability levels and test items aids in assessing the combination of ability grouping vs. items for which the model had the most difficulty estimating.

Patterns of the size of the residuals against the classical item discrimination indices reveal whether the assumption of equal discrimination indices is valid. If a plot of the absolute values of the standardized residuals vs.

the classical item point-biserial correlations reveal a curvilinear relationship, this is an indication that items with high and low biserial correlations had the largest residuals. A comparison of plots for the one-, two- and three-parameter models against classical item discrimination indices are expected to reveal better fits are obtained when the discrimination powers of test items are allowed to vary. To further examine the relationship between the size of the standardized residual to the discrimination indices, for each model, the absolute-value of the standardized residuals and item discrimination indices are categorized into groups and a crosstabulation is expected to reveal the fit to the data will increase with increasing number of fitted parameters.

To determine the relationship between the standardized residuals and the level of classical item difficulty, the items are divided into those with p values less than or equal to .5 and those greater than or equal to .5. Using absolute values, the standardized residuals are classified into those less than or equal to one and those greater than one. For the one-, two- and three- parameter models, a crosstabulation of the number of cases falling into each category is expected to reveal patterns of fit based on the difficulty of the item. For hard items, the three-parameter model is expected to account for the greatest percentage of residuals less than one. This is an indication that the three-parameter model, in its accountancy for guessing, has provided the best estimates of performance. The inference will be that guessing was a factor with the hard items.

For the one-, two- and three- parameter logistic models, a crosstabulation of the average and absolute average raw and standardized residuals across the

twelve ability categories is expected to reveal the three parameter model to be a better fit to the test data than the one- or two- parameter logistic models. To determine why some items fit particular models, the items are separated into four categories: those which fit the one- parameter model, those which fit the two-parameter model, those which fit the three- parameter model, and those which had similar fit across all three models. The characteristics of misfit are assessed by determining whether the item is discriminating or nondiscriminating, easy or hard, whether guessing might be a factor and for which ability levels the item appear to function.

For each item, plots of the standardized residuals versus ability are obtained for the one-, two- and three parameter models. Items with standardized residuals centered about zero indicate the model fits the test dataset. In comparison to the one- and two-parameter models, per item, the three-parameter model is expected to exhibit the best patterns of fit.